INTEGRATING MANUAL AND AUTOMATED METABOLIC ENGINEERING METHODS

BY

MATTHEW AARON RICHARDS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Chemical Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Professor Nathan D. Price, Chair
Professor Christopher Rao
Professor Huimin Zhao
Professor William Metcalf

Abstract

Metabolic engineering—the process of altering an organism's metabolism to achieve a desired goal—presents an alternative to established chemical processes. By tapping into the enormous range of metabolic transformations carried out by microbes, we can create microbial cell factories that improve upon chemical processes efficiencies while eliminating potential waste products. One of the greatest obstacles to realizing this potential is our incomplete picture of the possibilities of metabolism, a challenge we address by developing automated tools and databases to guide our inquiries. Though these high throughput studies have greatly accelerated the timeline to go from sequencing a genome to piecing together overall metabolic functions, these accelerated results generally come at the price of higher error rates. In my thesis work, I investigated ways to mitigate this loss of precision by more deeply integrating manually curated information into automated approaches.

In the first part of my thesis, I focused on defined microbial growth media, the essential substances that comprise the raw materials of biochemical processes. The vast majority of microbes cannot currently be cultured in a laboratory, a formidable obstacle to characterizing these organisms and their metabolisms. Methods that predict new defined media could expedite culturing experiments; however, such efforts require a repository of known defined media that collects successful growth conditions. To address this need, I created MediaDB, an open access database of chemically defined microbial media from published biochemical literature. MediaDB enables studies across different media that can reveal emergent trends in known media formulations across organisms. By examining media in the database, I found that they often contain similar trace mineral and vitamin solutions, suggesting a measure of uniformity in the way that biologists have traditionally created growth media. Clustering organisms based on their media compounds, I found no connection between media similarity and organism phylogeny, though several cases demonstrated a link connecting media to specific metabolic functions.

For the second part of my thesis, I built a genome scale metabolic reconstruction for *Methanococcus maripaludis,* an archaeon that produces methane from CO_2 and H_2. *M. maripaludis* provided an excellent engineering target, both for modifying forward methanogenesis as well as for working to oxidize methane to methanol, a first step towards building a pathway to liquid fuel that is of interest to the Department of Energy. I reconstructed my metabolic network model by relying chiefly on manual, resulting in the first network to correctly depict hydrogenotrophic methanogenesis. My reconstruction demonstrates the importance of electron bifurcation in central metabolism, providing both a window into hydrogenotrophic methanogenesis and platform to generate metabolic engineering hypotheses. I validated my model on growth yield and gene knockout data, showing its strong ability to reproduce experimentally measured results. Using the completed network, I predicted the previously unknown gene for glycine biosynthesis, a hypothesis I am now verifying with auxotrophic growth experiments. Moreover, I generated strain designs to achieve energetically feasible conversion of methane to methanol and in doing so, further demonstrated the vital role of manual curation for these predicted engineering strategies.

For the final piece of my thesis, I explored how to leverage manual curation to improve automated metabolic reconstruction. To this end, I created a method that “morphs” a manually curated metabolic model to a draft model of a closely related organism. My method combines genes from the original manually curated model with genes from an annotation database to create a final structure that contains gene-associated reactions from both sources. I used this method to create morphed models of three methanogens from iMR540 and showed that phylogenetic similarity between the source and target organisms correlated with the similarity of their models. I also found that gene annotations from iMR540 showed very low intersection with those from the annotation database, demonstrating the volume of information added by my manual curation. The morphing method could provide a viable

alternative to other automated reconstruction methods for organisms that are dissimilar from those that form the foundation of annotation databases.

Together, my work exemplifies the advantages conferred by integrating manual methods with automated tools. My studies demonstrate the importance of maximizing the information we glean from manually curated data and blending that data with automated tools that accelerate large scale studies of metabolism. Such approaches mitigate the pitfalls associated with relying solely on automated methods, ensuring the high quality and depth of data as we work to characterize the space of microbial metabolism.

Acknowledgments

This is easily my most anticipated section of this thesis document because I am eager to recognize the enormous support I have been lucky enough to receive over the past 5+ years. I have been fortunate to work with a remarkable group of individuals and to spend my non-work hours with a fantastic collection of friends and family. It is difficult for me to express in words what the support of all of you has meant to me as I've experienced the ups and downs of graduate school, but I will attempt to do so anyway. I tried to promise myself that this section would be brief, but I know myself well enough to admit that brevity is not my strength.

First, I would like to thank Nathan Price for taking me on as a graduate student and advising me during my studies. You not only brought me into the group but also brought me out to Seattle and have provided me with numerous opportunities for personal and professional growth. Thank you to Shuyi Ma (tra la la) for being my first friend in the lab, for routinely being my lunch buddy at ISB, and for patiently listening to so many of my ideas, even the crazy ones. Thank you to Matt Benedict for teaching me about methanogenesis and modeling and for being one of the flat out nicest people I've ever met. Thank you to Nick Chia for being my first post doc mentor, for pushing me in the right direction during my early graduate school years and keeping up that contact even after leaving the group. Thank you to Ben Heavner, the most thoughtful scientist I know, for teaching me so much about the right way to conduct research, for mentoring me during my time in Seattle, for helping me improve my writing skills, and for shaping my ability to think scientifically. Thank you to Vangelis Simeonidis for pushing me to finish my first paper when I thought all was lost, for selflessly volunteering to help me as I assembled this thesis, and for helping me to develop my own confidence as a researcher. Thank you to Roie Levy for insisting on helping me to think through numerous ideas in the last year and for tirelessly reading through some of these chapters to make sure my head was screwed on right. Thank you to James Eddy for numerous

nuggets of wisdom over the years, for helping me immensely to transition to Seattle in 2012, and for introducing me to so many fantastic things to pursue during my off hours. Thank you to Saheed Imam for providing so much modeling wisdom in such brief time and for tolerating all my soccer-related taunting. Thank you to Sascha Schauble for all of your advice, for convincing me of my capabilities, for last minute figure alterations, and for being such a wonderful friend. Thank you to Victor Cassen for months of working with me to make my website a reality, including far too much time spent revising my scripts. Thank you to Ram Hariharan for keeping me on my toes, for reminding me of the virtues of Python, and for letting me follow him from ISB to UW and back again. Thank you to Caroline Porter for encouraging me to complete my preliminary exam and coaching me through the process. Thank you to Ali Paquette for constantly reminding me that I belong here, for being understanding of my anxious moments, for exemplifying the hard work and dedication needed to accomplish great things, and, above all, for being my good friend and confidant. And to everyone else in the Price Lab today—Seth Ament, Dani Bergey, John Earls, Cory Funk, Piyush Labhsetwar, Hongdong Li, Daniel McDonald, Jocelynn Pearl, Vineet Sangar, Paul Shannon—as well as many others who have come and gone in the group, thank you so much for your support and encouragement. It has made such a positive impact on me to work with all of you, I have immense respect for every person in the lab and I can't imagine working with a better group.

So much of my work would not have been possible without a remarkable group of undergraduates who I mentored during my time in graduate school. My MediaDB group—Hao Feng, Zhilong Zhu, Colin Hoffman, Nassim Ajami, Andrea Novitsky—tolerated their first year mentor extremely well and I appreciate their patience as I figured out how to do research correctly. A special thanks to Nassim and Andrea who contributed so much time to the database and were instrumental to making it a reality. Thank you to Vy Nguyen, my fantastic mentee in 2013, for so many conversations about life, science,

food, and happiness. I would not be the teacher I am today if not for lucking into having you as my intern that summer. And a huge thank you to Brendan King, my summer intern and current undergraduate researcher who has brought so much programming knowhow to the model morphing project. You have continually impressed me with your capability to learn about metabolic modeling while also navigating the quirks of Kbase and I have no doubt that you will go on to do great things.

I want to give extra special recognition to Theresa Fitzgerald, who I sincerely believe is the glue that holds the lab together. You have been there for me time and time again, helping me hold myself together during times of frustration, anxiety, sadness, and anger while somehow managing to do all the administrative work behind the scenes. Your cheery demeanor is one of my favorite parts of coming to work and I cannot thank you enough for your friendship and your help over the past 4 years.

The final 1.5 years of my PhD have been enriched by working with the Leigh Lab, a wonderful microbiology group at the University of Washington. I want to thank John Leigh for allowing me to come in and do experimental work, for welcoming me into your group, and for so many good discussions about biochemistry, academia, and life. Thank you to Thomas Lie for teaching me so many things, for being patient with my limited knowledge of general biology, for tolerating the full radiance of my large personality, for being a superb mentor with lots of good advice, and for doing all of that with a perpetual smile on your face. Thank you to Eli Gachelet, my partner in crime in the wet lab, for being my cloning guru, for many conversations about every topic, for helping me invent and perpetuate all sorts of ridiculous ideas, and for brightening my cloudy days. Thank you to Chloe Hart for all the sarcasm, all the Tom-foolery (Ha!), for absurd banana dances, and for many talks about graduate school and games. Thank you to the remaining lab members for your help with experiments and for welcoming me so readily into the lab.

Outside of lab work, I have been profoundly affected by my friends at the University of Illinois and ISB. I want to thank my friends Cartney Smith, Jessica Banks, Kathryn Trenshaw and Mike van Wren for so many good nights filled with bad movies, video games, made-up stories full of silliness, and tasty food. Thank you to Kristine Pangan-Okimoto for introducing me to yoga and being my workout buddy, years after sitting together in middle school PE class. Thank you to Nik Dudukovic who dove into underwater hockey with me to form the dynamic duo; you are the best teammate I have ever had. Despite what we told people about the nature of our dynamic duo, I will always believe that you are the Batman to my Robin. All of you that I have mentioned here and a large number of others were my favorite part about being in Champaign-Urbana; you made the first two years of graduate school an unforgettable experience. Choosing to move away was among the hardest things I have ever done and it's because you were all so wonderful. I regret that my journey to this point had to take me away from all of you, but I fully expect to reconnect many times in the future.

In Seattle I want to thank all of my fantastic friends; there really are far too many to name them all but I will try my best. Thank you to Hannah Cox who vowed to drag me through my PhD kicking and screaming if need be and who I miss dearly since she left ISB. Thank you to Aaron Brooks for all the trips to the climbing gym, all the games of foosball with the rest of the Baliga Lab gang, and conversations about the ins and outs of grad student life. Thank you to Nina Arens for giving me the opportunity to do more science education outreach and for bringing so many conversations about chocolate into my life. Thank you to Cora Chadick for teaching me your silly fighting technique, for being unafraid to try out my sily sport, and for so many ridiculous conversations that enriched my life. Thank you to Karlyn Beer for giving me silly nicknames and splitting apple fritters with me on Top Pot day. Thank you to Cameron Conway for being there for me during my first year in Seattle and listening to me, even when you were the one who really needed an ear. Thank you to Christina Jones for being my Seattle underwater hockey

best friend, to Paul Byrne and Pat Carboneau for welcoming me emphatically to the Seattle team, and to the rest of the Seattle Seahammers for all the great hockey. Thank you to Sarah Present for embracing my craziness and balancing it with so many meaningful conversations about life; I will always be grateful that we both spontaneously went to a random meetup in Bellevue. Thank you too all my friends from the Jewish community, including all of my co-travelers on the 2014 Birthright trip. Thank you to Kyle Caldwell and Paula Olson for years of friendship and for spending virtually every weekend playing games, eating cheese, going on trips, and doing everything else in between. Thank you to Rosemary Royce for welcoming me readily into your family, for understanding my moodiness, and for always thinking the best of me and my abilities. Thank you to Robin McCoy Brooks for countless walks around Green Lake that became essential to my mental well being, for teaching me about the pleasures of experiencing the world around me, and for never being afraid to speak your mind.

I want to profoundly thank my family for being there for me and supporting me all these years. To my grandmother, aunts, uncle, and cousins—Helene, Nancy, Robin, Doug, Amanda, Jenessa, Brynnea, Cameron—I want to thank you for always having my back, for all the years of going out of your way to spend time together, and for helping shape th person I have become. To my siblings, Ben and Rebecca, thank you for your patience with your big brother, for all the conversations about life, and for sharing all the important moments in our lives with one another. And to my parents, Dean and Andrea, thank you for illuminating the path I have followed, for never giving up on me, for insisting that I am capable of excellence, and for all the values you have instilled in me over the years.

To Molly Long, who has stuck with me through 3 years of ups and downs, thank you for being my companion, for listening to all of my thoughts, for always insisting that I can achieve my goals, for laughing through the good and crying through the bad, and for selflessly putting my needs above your

own. I cannot imagine the last few years without you, nor would I care to do so. You have been the constant throughout my tumultuous path to this point and I can never thank you enough for that.

Finally, I know that when I began graduate school I had 4 living grandparents, a number that has sadly dwindled to just 1. To my paternal grandparents, Darrel and Maxine, thank you for all the trips to Iowa, for all the cookies and applesauce, for all the cold mornings in the woodshop, and for every little quirk that made those trips one of my favorite parts of growing up. To my maternal grandfather, Myer, thank you for all the years of coming to all of my events, for teaching me how to be a *mensch*, for being so open to embracing my intellectual and culinary pursuits, and for entrusting me to carry on your name. Though I miss you terribly and wish you could have shared this accomplishment with me, I can never regret the time we spent together.

Table of Contents

Chapter 1: Introduction 1

Chapter 2: MediaDB: a database of microbial growth conditions in defined media 18

Chapter 3: Exploring Hydrogenotrophic Methanogenesis: A Genome Scale Metabolic Reconstruction of *Methanococcus maripaludis S2* 38

Chapter 4: Guiding Strain Design with the iMR540 Metabolic Reconstruction 70

Chapter 5: A Method for Perpetuating a Metabolic Reconstruction 95

Chapter 6: Discussion and Conclusions 121

Chapter 7: References 132

Appendix A: Supporting Information for Chapter 2 147

Appendix B: Supporting Information for Chapter 3 160

Chapter 1: Introduction

DNA Sequencing and Metabolism

Biology is well entrenched in its information age, with more tools available than ever before to help us understand the complexity of living systems. Perhaps the most ubiquitous example of this age is the evolution of DNA sequencing technology. A useful beginning reference is the advent of Sanger chain elongation sequencing in 1977 [1], the first technology to enable manual DNA sequencing in a non-destructive way. This initial technology became the backbone of the myriad of technologies that followed, innovations that massively increased the speed at which we can sequence full organisms and survey metagenomes, enabling many more technologies that leverage sequencing data. Since the first fully-sequenced genome for bacteriophage ϕX174 was published 38 years ago [2], our collection of complete genome sequences has grown to encompass over 58,000 species per the NCBI RefSeq database [3]. Such an abundance of genomic information presents a plethora of opportunities to use the available information for performing both isolate and interspecies studies.

GEnome scale metabolic Network REconstructions (GENREs) present a promising avenue for using sequencing information to elucidate mechanisms that drive organisms' metabolism. A GENRE is built on top of gene-protein-reaction (GPR) relationships that link the genome to metabolism[4]. In these constructs, a complete organism's genome is converted into proteins and annotated by mapping each protein or protein complex to the reaction(s) it catalyzes. The stoichiometry of these reactions is represented in a stoichiometric matrix (S-matrix) that ties each reaction to its metabolites using stoichiometric coefficients. The resulting reaction network depicts, to the fullest extent of our knowledge, the scope of metabolic reactions occurring within the sequenced organism, essentially

ascribing functional roles to all metabolic genes. A reconstruction provides a powerful way to look holistically at metabolism and to analyze the various mechanisms at play. As such, it can also serve as a platform for generating strain design hypotheses, making a reconstruction a valuable tool for metabolic engineering efforts as well [5].

Genome scale metabolic reconstructions form the basis for much of the work performed for this dissertation, thus many introductory concepts regarding their nature are explained in the following chapters. However, given the volume written regarding metabolic reconstructions in following sections, a brief word on some nomenclature is in order, specifically regarding the distinction between a metabolic reconstruction and a GEnome scale Model (GEM). An excellent overview of the differences between these two constructs is described by Heavner et al [6]; for the purposes of this work, they can all be distilled to the ability to simulate growth. A GENRE is a network for organizing known metabolic information for an organism, thus it can be manually examined but cannot simulate growth. By contrast, a metabolic model is a mathematical construct and thus must be able to simulate growth. For mathematical completeness, a GEM must necessarily contain complete synthesis pathways for all components of cell mass; hence it must fill any network "gaps" that would otherwise prevent growth. Accordingly, a GEM supplements the established information in a GENRE with potentially hypothetical additional reactions to achieve mathematical completeness, a process known as gap filling. This major inherent difference manifests in other diverging features between GENREs and GEMs, as shown in Figure 1.1.

Among these differences, the critical elements of GEMs all relate to subjecting the mathematical model to constraints that reduce the size of the model's solution space. For example, within the map of a reconstructed network, metabolic reactions can be assumed to carry any magnitude of reaction fluxes; reactions in a GEM are much more restricted and are often constrained to carry only a certain level of

flux. Thus, in the chapters that follow I often use the terms “constraint based modeling” and “metabolic modeling” interchangeably. This is distinct from my use of the terms “reconstruction” and “model”, which are inherently different. All that being said, it is nearly impossible to discuss one construct without at least acknowledging the other. Reconstructions and models are intrinsically linked, with the former providing a scaffold that directly leads to the latter and the latter providing predictive functionality that utilizes the information in the former. Thus, any sort of tool or innovation that affects a GENRE must also impact a GEM and *vice versa*. Having digressed sufficiently at this point, let us return to the matter at hand, which is that of metabolic networks and how they are analogous to DNA sequencing.

An Analogy Between DNA Sequencing and Metabolic Reconstructions

Strikingly, the evolution of genome-scale metabolic reconstructions strongly parallels that of DNA sequencing itself (reviewed by Mardis [7]), albeit over two decades later (see Figure 1.2). Much like modern DNA sequencing technologies can logically be traced back to Sanger’s sequencing in of phage φX174 in 1977, genome scale metabolic network reconstruction can be traced to the genome-scale network of *H. influenzae* published in 1999 [8]. This first GENRE laid the groundwork for all reconstructions that followed it, sketching out the necessary steps and demonstrating the utility of the final product. Both technologies were completely manual in their inceptions, requiring months of dedicated effort to produce a final structure. Unsurprisingly, both of these examples contained multiple manual bottlenecks that presented opportunities for improvement through standardization and automation. In the case of DNA sequencing, running fragments in a gel, exposing them to X-rays, and reading the results by eye were all labor-intensive manual steps with much room for improvement [7]. As for reconstructions, much effort in the manual protocol was dedicated to tracing each individual gene to its proposed metabolic function through arduous literature searches [4]. Recognizing the abundance

of manual bottlenecks in both processes, the scientific community responded by developing numerous improvements for these first efforts, revolving around introducing tools to automate much of each procedure.

DNA sequencing received perhaps its most important upgrade in 1986, with the commercialization of fluorescently-labeled automatic sequencing [9]. This automated sequencing platform removed many manual steps required to process and read the resulting sequences, reducing some previous sources of human error and enabling much more ambitious sequencing efforts. It essentially paved the way for realistically sequencing whole genomes in multicellular organisms, including the beginning of the Human Genome Project. A parallel development for metabolic reconstructions was the birth of annotation databases, arguably the most crucial element for quickening the pace of metabolic reconstruction. An annotation database is simply a collection of manually-annotated genes and provides a repository of proposed gene functions (genes and their GPRs) based upon previous biochemical characterizations. The first major annotation database was the Kyoto Encyclopedia of Genes and Genomes (KEGG), published in 2000 [10] and followed by MetaCyc in 2004[11]. Such resources were hugely important because they centralized much of the information needed for a high quality reconstruction, eliminating a large amount of literature search time. Furthermore, databases established standard identifiers for metabolites and reactions, essentially introducing languages with which to build and share reconstructions. Such resources represented an important step forward in speed akin to the advent of the automatic DNA sequencer for DNA sequencing, making it much more realistic to reconstruct networks, even an ambitious reconstruction of human metabolism [12]. In both cases, the underlying procedure remained largely the same, but the existence of a central tool—either a machine sequencer or central database—significantly sped up the process. As a result, automated Sanger sequencing with

fluorescent labeling dominated the DNA sequencing landscape up until about a decade ago [13] and nearly every GEM and GENRE published in the last decade stems from an annotation database.

In the case of DNA sequencing, the next major process improvement was that of the capillary sequencing instrument 1997 [14]. By employing fixed capillaries, this method eliminated slab gels in the sequencing process, alleviating both the time and effort needed to prepare, dry, and read a gel. The subsequent gains in speed from this method enabled major undertakings, giving rise to many of the reference genomes we now possess [7]. Somewhat analogously, metabolic reconstructions got a technological boost in 2007 with the publication of the GapFind and GapFill algorithms, automated methods that accelerated the gap filling process needed to convert a GENRE to a GEM [15]. These methods, which filled network gaps using a maximum parsimony (*i.e.* "shortest path") approach, removed virtually all of the manual work necessary for taking a metabolic network and finding ways to make it predict growth. Much as capillary sequencing enabled full genome sequencing for many model organisms, automated gap filling greatly enhanced our ability to create fully functional GEMs with little effort. Though both innovations still contained the same basic set of steps, they eliminated more manual bottlenecks, greatly speeding the overall processes of DNA sequencing and metabolic modeling, respectively.

As reviewed by Shendure and Ji [16], DNA sequencing progressed into a "second generation" or "next-generation" of techniques beginning in 2005. That year the 454 pyrosequencing platform became commercially available [17] and was followed in the next few years by a variety of other techniques, such as the Solexa platform based upon bridge PCR and reversible terminators [18,19] and the HeliScope platform based on single molecule techniques[20,21]. The broad group of second generation techniques is linked by their parallelization of sequencing and detection steps; hence they enabled what is commonly referred to as "massively parallel sequencing" [7,16]. This parallelization has made possible a

complete paradigm shift in the way that we regard DNA sequencing, with sequencing reads shrinking in individual size but multiplying many times in number [16]. In short, next generation sequencing has completely changed the way we sequence and as a result, the last decade has witnessed easily the most rapid technological growth for DNA sequencing in terms of sheer number of advancements. A similar change took place for metabolic reconstructions and models in 2010 with the publication of the Model SEED automated reconstruction algorithm [22]. In using the SEED database of genomes and annotations to build a metabolic network directly from a genome, this algorithm mostly eliminated the manual steps of pulling reactions from a database. Analogously to DNA sequencing, the Model SEED method performed multiple steps, annotating a supplied genome to create a reconstruction and using automated gap filling to create a model capable of simulating growth. Like next generation sequencing methods, this new innovation changed the landscape of metabolic reconstructions, enabling high throughput generation of a functioning GEM for any organism with a completely sequenced genome. Moreover, by creating a more user-friendly set of tools for building reconstructions or simulating models through a web interface, the Model SEED made metabolic reconstruction and modeling more widely accessible to the biological community, just as decreased costs and processing time associated with next generation sequencing have made DNA sequencing accessible for most research groups.

These parallel paradigm shifts have not come without consequence; for their prodigious gains in speed, both DNA sequencing and metabolic modeling have sacrificed a measure of depth to achieve wider breadth. In next generation sequencing, relying on many short reads has increased speed dramatically, resulting in a larger total pool of information for the sequenced organism but less sense of how the reads fit together (*i.e.* lower raw accuracy when practiced in *de novo* fashion) [16]. For this reason, large sequencing efforts typically eschew less accurate *de novo* sequencing in favor of mapping reads onto a reference genome assembled using more accurate but slower Sanger techniques [7]. Similarly, although

automated reconstruction has enabled us to create models for many more organisms, these models lack organism-specific richness and run the risk of resembling the models of more common organisms. This is illustrated by the fact despite increasing publishing rates for reconstructions, recent GENREs tend to overwhelmingly contain the same reactions as previously published models rather than adding their own unique reactions [23]. To mitigate the effects of incorporating the same generic reactions into each model, these automatically-generated models must be treated as first drafts; expanding and enriching these models relies on literature searches and biochemical characterizations.

Of course, despite their similarities, DNA sequencing and metabolic reconstructions differ considerably in terms of their current capabilities. Likely due to the fact that sequencing predates genome-scale reconstructions by over 2 decades, existing technologies for metabolic reconstructions lag behind those for DNA sequencing. DNA sequencing has arguably progressed on to a third generation of techniques that leverage single molecule sequencing to achieve even greater gains in speed and cost efficiency [24]. As a result, although the number of metabolic reconstructions has grown substantially in recent years, the several hundred networks in existence pale in comparison to the 58,000+ sequenced genomes in the RefSeq database. Though closing this gap is improbable based on the ever-increasing speed of sequencing and impossible due to a GENRE's inherent dependence on an existing DNA sequence, there appears to be ample opportunity to increase the speed of reconstruction with more advances akin to those already realized in sequencing. However, working toward this end will require that the metabolic reconstruction community begin to address a number of challenges that face GENREs.

Challenges for metabolic reconstructions

Arguably the greatest obstacle to creating high-quality reconstructions is the relative dearth of available biochemistry information. Considering the incredible breadth of microbial species, the bulk of known

biochemical phenomena have been characterized in a miniscule number of organisms. For example, compared to the 58,000+ complete genomes in RefSeq, the MetaCyc database has pathway databases for 7,668 organisms, or about 13% of fully sequenced organisms [25]. The remaining majority of organisms occupying the rest of the sequenced microbial biosphere represent a huge space of unknown function, greatly limiting our ability to construct reaction networks for these organisms. Even within the well-characterized organisms, we still annotate many genes as “hypothetical proteins”, enzymes with no known function that may presumably participate in metabolism. With respect to the overwhelming volume of unknown information, these unknown portions of metabolism ultimately limit the predictive power of GENREs and GEMs, decreasing the degree with which we can use such constructs to discover emergent network properties. However, this same limitation presents an opportunity to use metabolic reconstructions to propose functions for unknown portions of metabolism, particularly by simulation as GEMs.

Another obstacle to metabolic reconstructions is the generic nature of annotation databases. Like next generation sequencing, which trades increased speed and volume of reads for decreased overall read precision, annotation databases speed up the reconstruction process but are also fairly error prone. Although in both cases this tradeoff results in a much larger overall breadth of data than could be achieved with fewer, more precise pieces of information, this advantage is greatly reduced for GENREs because it generally ends up requiring many additional manual corrections [26]. Annotations derived from such databases come from gene homology and are not necessarily even consistent across databases. An illustrative example of this problem is found in the reconstruction of *C. beijerinckii*, which was derived using annotations from three separate annotation databases [27]. As demonstrated by the authors, very few reactions in the final reconstruction appeared in all 3 databases and many appeared in

only one database. Thus, any GENRE built using an annotation database may contain numerous mis-annotations and will necessarily be biased toward the annotations in that database.

Annotation errors and differences between databases are further complicated by the lack of standardized reconstruction format and nomenclature. A strength of DNA sequencing is its integration into standard file formats—FASTA [28], GenBank [29]—and a standard alphabet—A, G, T, C—such that any publically available sequence can easily be used by any researcher in the world. Although several groups have attempted to impose or suggest standards for biochemical networks, both generally for all systems [30] and specifically for metabolic networks [31], the current scope of available GENREs exist in a variety of formats and identifiers. This lack of extensibility presents a major problem, as many published GENRE and GEMs reportedly cannot be used to reproduce results in from their reference materials [31,32]. Furthermore, disparate nomenclature greatly complicates efforts to update and consolidate multiple GENREs of the same organism into consensus networks, as in e.g. the first consensus yeast network [33]. This particular complication is exacerbated by the large number of GENREs that lack any standard identifiers [31]. The issue of a standardized format has been addressed to some extent by the adoption of the systems biology markup language (SBML) [34], a format for representing all biochemical networks that has recently been updated specifically for constraint-based networks [35]. However, the metabolic reconstruction community as a whole has yet to unite itself behind any particular standard and thus, GENREs still exist in a variety of different formats [31].

Standardization problems are further magnified when considering the lack of a central repository specific to constraint-based network reconstructions. To once again borrow from DNA sequencing, genome sequences can be deposited straightforwardly in the RefSeq database by any group in the world, provided that they adhere to the standards upheld by NCBI [36]. Thus, RefSeq exists not only to centralize all complete genome sequences in one hub, but also to enforce formatting and nomenclature

standards on its content. Furthermore, the NCBI interface intrinsically houses genome sequences in a framework that permits usage of common analysis tools, such as BLAST+ [37]. At this time, no comparable database exists for constraint based networks; rather, the modeling community maintains several disparate resources aimed at serving different functions.

Table 1.1 displays a set of desired database functions, with performance of current databases as compared to these guidelines; as demonstrated, no one database contains all of these functions. The BioModels database [38] is designed as general biochemical model repository that adheres to established MIRIAM standards [30] and contains many manually-created models and networks. Thus it lacks any additional tools and is not specific to constraint-based models or networks. The BiGG database 2.0 [39] is specifically for constraint-based reconstructions and allows users to download GENREs or examine reaction pathways through its web interface. However, it is primarily an outward-facing resource, designed to facilitate use of the GENREs from one particular lab group rather than all known reconstructions. Kbase, built on top of the Model SEED [22], fulfills the largest number of the guidelines specified in Table 1.1 because it allows upload of new networks and provides tools for building and downloading new networks and models. Yet Kbase also falls short of meeting these standards as it does not contain an outward-facing database of manual reconstructions and does not rigorously enforce formatting or nomenclature standards. Examining these resources, it is evident that the metabolic modeling community currently lacks a unified standard of how reconstructions are created, stored, and distributed, a hurdle that must be cleared if GENREs are to ever achieve the current usability of DNA sequences.

Manual and automated: a combined approach

In examining these four primary challenges, it is useful to segregate them into two separate groups: those that affect individual reconstructions and those that affect the entire reconstruction community.

The latter two obstacles—need for nomenclature/format standardization and lack of a centralized repository—present significant difficulties for reconstruction and model usability, particularly in reconciling and comparing networks constructed by different groups. Resolving these problems will likely take a widespread commitment to adopt universal standards and designate a standardized central database for storing and using GENREs. Until such steps are taken, it is vital that any published reconstruction be explicit in explaining all of its facets and subjective annotation decisions to enable better sharing and usability [40].

However, the former two obstacles—lack of sufficient data and abundance of mis-annotations in existing resources—present obstacles to producing a high-quality reconstructions. Remarkably, these obstacles bear a resemblance to current challenges inherent in DNA sequencing. Despite the speed gains in both processes resulting from increased automation and technological advancement, both DNA sequencing and metabolic reconstruction are still somewhat constrained by error. Much as decreased sequencing read lengths have magnified the influence of sequencing noise and error, metabolic reconstruction's increased reliance on centralized databases and high throughput algorithms has exacerbated the problem of missing or incorrect gene annotations. The solution to both of these problems lies in the ability to compare to high quality reference information; in the case of DNA sequencing, this involves comparing a new sequence to a reference genome like those produced using capillary sequencing. For metabolic reconstructions, this means returning to high confidence information from biochemical literature sources; in short, it necessitates employing meticulous manual methods.

On its surface, relying on the manual-based curation is somewhat counterintuitive as it seemingly represents a step backward in the evolution of reconstruction techniques. However, manual curation is vital for producing a high-quality reconstruction. Any automatically-curated reconstruction can only be

considered a first draft and is generally followed by 6-12 months of manual effort [4]. This consists of sifting through virtually all available biochemical literature and sometimes multiple annotation databases to uncover the maximum amount of biochemical phenomena occurring within an organism. In putting forth this effort, a metabolic modeler can alleviate the detrimental effects of mis-annotations by replacing incorrect GPRs with correct annotations from literature sources. Furthermore, even though overall data limitations are still difficult to overcome, a high-quality manual reconstruction can shed light on many areas of metabolism and point directly at portions of pathways that still require biochemical characterization. Thus, although automated methods enabled by annotation databases and algorithm development have shortened the timetable needed to produce a high-quality reconstruction, the current state-of-the-art process is a combination of these automated resources and careful manual curation.

Manually curated GENREs are but one example of the current duality present in biology, where pervasive manual techniques are employed in concert with automated methods. Routine tasks within the field are being increasingly automated to cut down on processing time by leveraging technological advances and growing computational power. With the rise of computational biology, an increasing amount of data analyses are performed *in silico* on large datasets, creating a pipeline for quickly extracting meaningful information from experimental datasets. But ultimately, these tools often produce a starting point for using manual techniques, the gold standard for achieving maximum accuracy. Even with the huge speed gains we have experienced in recent years, current computational models are often inadequate for solving complex biological problems without substantial manual effort. Thus, procedures that blend automated and manual methods hold considerable promise for addressing a variety of problems.

A hybrid approach to developing growth media

As I will describe in Chapter 2, integrating manually-curated data into computational methods could greatly aid the field of microbial growth media design. Traditionally, new defined growth media are developed by manually adapting existing media to fit the needs of a new organism. The lengthy, arduous nature of this manual approach is well illustrated by the SAR11 clan, perhaps the most ubiquitous group of organisms in the world [41]. After discovering these organisms in 1990, researchers required another 12 years of work to first isolate a culture in complex media [42,43]; developing a working chemically defined medium took another 11 years [44]. With the growing wealth of available sequencing data for uncultured organisms, it is possible linking organisms' genomes to their growth media might hold some sort of key for speeding up this process [45].

This concept laid the groundwork for MediaDB [46], the first publically-available repository of defined microbial growth media from literature sources. MediaDB, discussed further in Chapter 2, was followed by the KOMODO database [47], a larger repository that included an algorithm for predicting organism growth media based upon phylogeny. Both of these resources created centralized spaces for studying existing media and trends across organisms, plus the addition of a media prediction algorithm presented an automated process to supplement manual work. Furthermore, topological network analysis methods, such as computing organisms' seed sets to determine compounds that must come from the environment [48], have also shown promise for relating genomic information to nutritional requirements. These tools could potentially speed up media development by informing the choice of a starting set of media compounds. However, a multitude of other factors influence organism growth, such as the need for dilute metabolite concentrations or siderophores from other organisms [49,50]. Thus, advances in media formulations are still dependent on manual methods that incorporate these variables, which are not currently considered by an automated media prediction algorithm.

Leveraging manual efforts to improve automated methods

As demonstrated by the cases of metabolic reconstructions and media formulations, advances in computational techniques have improved our capabilities on multiple biological fronts, in large part by greatly increasing the speed of repetitive tasks. Yet ultimately, these technologies still cannot achieve the same results as slower, more meticulous manual methods. Much as media formulations ultimately rely on experimental methods, high quality metabolic reconstructions depend on careful manual curation, even when beginning from automatically-generated starting points. Though there may be a time in the near future when advanced computational models eliminate the need for more of these manual procedures, current technologies cannot approach the accuracy achieved through manual means, creating a potential choice between quick results and accurate ones. Furthermore, it is not necessarily reasonable to expect that future technologies will completely eliminate the need for rigorous manual methods.

It is therefore crucial to improve the hybridization of manual and automated methods to better inform the high-throughput results of automated tools with meticulously gathered experimental data. Despite the fact that much of biology is information-limited, whether by inability to culture the vast majority of organisms or ignorance of uncharacterized biochemical pathways, there is a tremendous amount of manually-generated data available through biochemical literature. Current methods, such as those emerging in the space of developing microbial growth media, are beginning to integrate existing experimental data with more automated methods, but more work is needed to create effective workflows that maximally leverage manually-curated data. Through my dissertation work, I have strived to establish tools and methods that encourage increased usage of published biochemical knowledge, centered around the aforementioned issues of formulating growth media and building metabolic reconstructions. Specifically, my dissertation is structured as follows:

- Chapter 2 describes MediaDB, a manually-curated database of chemically defined microbial growth media. As the first resource of its kind, MediaDB provided a resource for using and studying known media formulations by bringing together growth data from many published journal articles into one central repository.
- Chapter 3 describes the creation of iMR540, a manually-curated metabolic reconstruction of *Methanococcus maripaludis* S2 built on top of automated reconstruction methods. In addition to providing a valuable tool for studying microbial methane production, this network aptly demonstrates the efficacy of manual reconstruction methods to leverage biochemical literature data for improving automated draft reconstructions.
- Chapter 4 describes multiple applications of the completed iMR540 reconstruction as a tool for studying methanogenesis, elucidating uncharacterized biosynthetic pathways, and generating metabolic engineering strategies.
- Chapter 5 describes a novel method designed to morph a manually-curated metabolic network into a functioning reconstruction of a novel organism. Unlike existing automated reconstruction, this method incorporates both automated gene annotations and manually-curated information from the original network, effectively using gene homology to create an automated draft reconstruction that incorporates much manual curation.
- Chapter 6 reflects upon the previous chapters and discusses the implications of this work as a whole, focusing on summarizing their contents and describing future directions of this work.

Together, these chapters will describe my efforts to integrate more manually gathered data into automated methods, creating processes that utilize the best qualities of both approaches.

Tables and Figures

Desired Database Feature	BioModels	BiGG 2.0	ModelSEED/ Kbase
Can upload a manually created GENRE/GEM	X		X
Checks for standards adherence and reproducibility	X		
Has tools for GENRE/GEM creation			X
Has tools for GENRE/GEM analysis		X	X
Allows GENRE/GEM downloads in standard format	X	X	
Specific to constraint based networks/models		X	X

Table 1.1: An evaluation of the 3 major metabolic reconstruction databases according to desired features

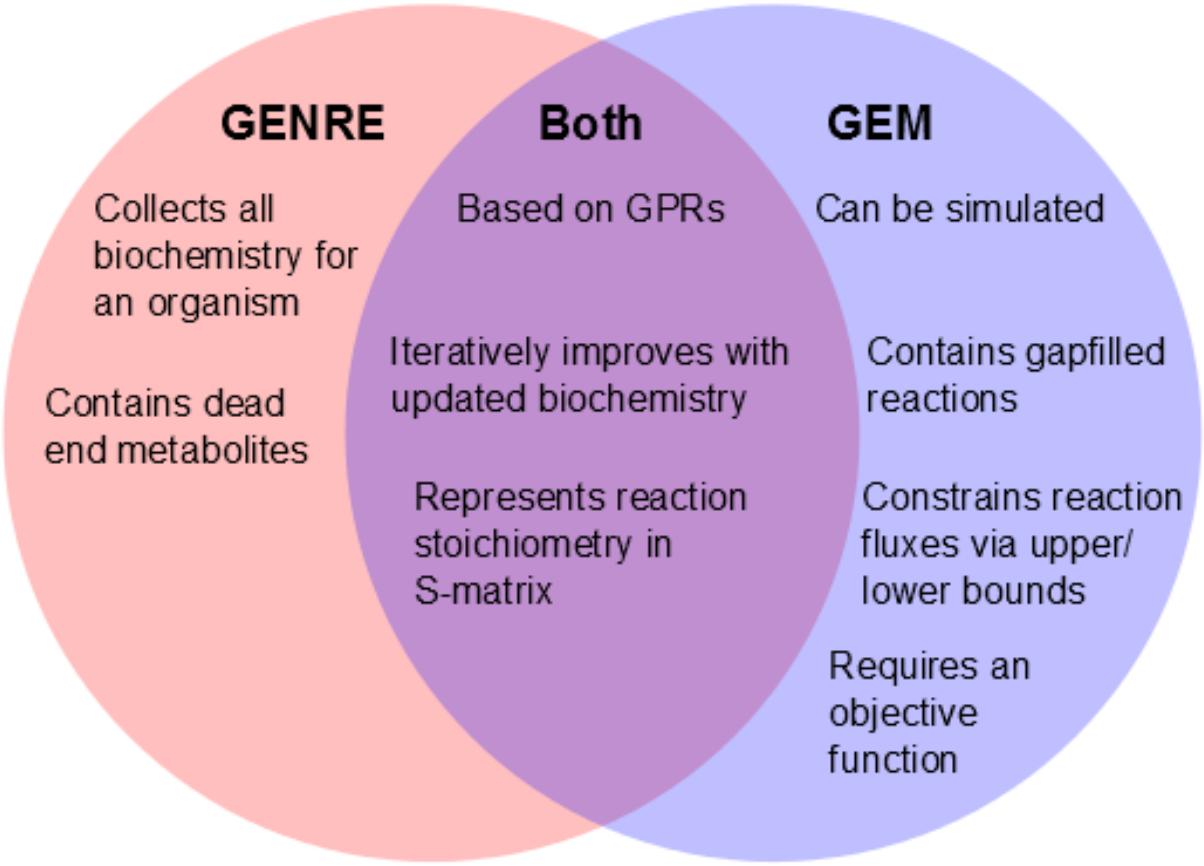

Figure 1.1: A comparison of unique and shared features between a GENRE and a GEM.

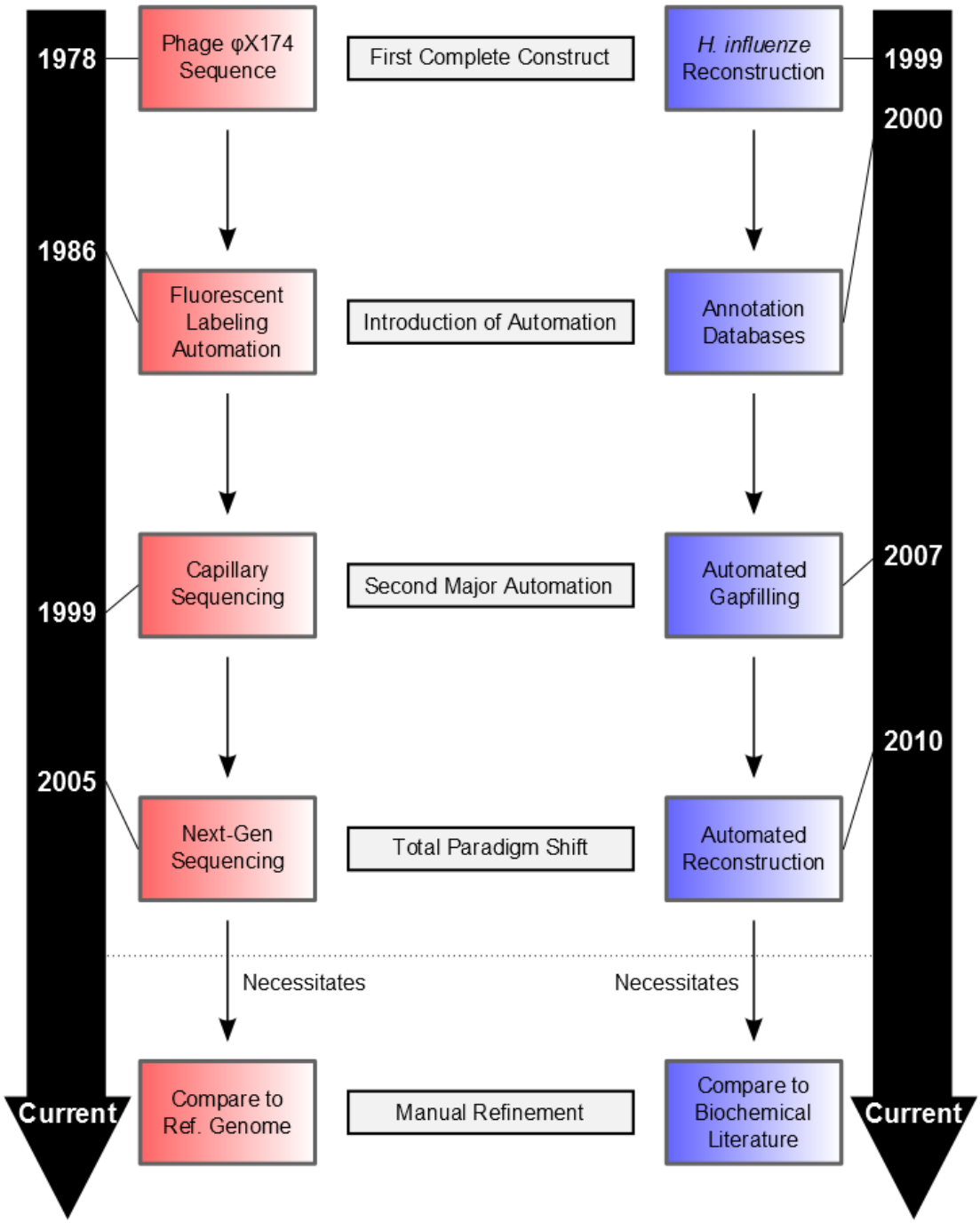

Figure 1.2: A side-by-side look at major advances in DNA sequencing (red) and metabolic reconstruction (blue). Text in gray boxes indicates the relationship that connects adjacent boxes; for instance, “Capillary Sequencing” and “Automated Gapfilling” represent the “Second Major Automation” in their respective fields.

Chapter 2: MediaDB: a database of microbial growth conditions in defined media[1]

Introduction

Genomic and high-throughput sequencing technologies enable the generation of large amounts of genetic information on microorganisms without the need to grow cultures in the lab. Armed with these technologies, we can automatically generate draft metabolic network reconstructions for organisms directly from genome annotations [51] and derive metabolic network models to simulate microbial growth *in silico.* These models can be improved through an iterative curation process between experimental and computational investigations [52]. To date, this iterative process has been most successfully advanced by partnering *in silico* reconstruction with *in vitro* characterization of isolates grown in defined laboratory media—an experimental approach that remains the most comprehensive method for characterizing microbial physiology [53–59]. Techniques for building metabolic network reconstructions from genomic data have progressed sufficiently to enable the application of *in silico* models for characterizing microbes that have not been cultivated *in vitro.*

Only 0.1-1% of the estimated number of microbial species have been isolated and successfully cultivated in a laboratory environment [55,56,58]. The collection of species we can currently culture spans only 30 of over 100 established phyla and mostly contains fast-growing organisms—organisms that are not the most prevalent species in the environment[57,59]. A

[1] This chapter is a reprint of a published article. The citation is as follows: Richards MA, Cassen V, Heavner BD, Ajami NE, Herrmann A, Simeonidis E, and Price ND. "MediaDB: a database of microbial growth conditions in defined media." *PLoS ONE* (2014) 9(8): e103458. doi: 10.1371/journal.pone.0103548

range of novel techniques have been applied in efforts to culture less characterized microbes, such as using diffusion chambers to mimic environmental conditions [60–63], adding growth factors or signaling compounds secreted from other organisms [64–67], diluting media nutrients to lower concentrations [43,49,68–72], increasing incubation time [69,70,72–76] and running high-throughput cultures [71,77–79]. These innovations have increased the diversity and number of culturable organisms, but the large number of factors that can affect *in vitro* growth still presents a challenge for isolating and culturing microbes from environmental samples.

Recently, computational modeling has been successfully applied to support culturing efforts. Several groups have used metabolic reconstructions, which are based on organism-specific genome sequence and biochemical knowledge, to assist in media design. Applications of these networks to media design have included both direct querying of the metabolic network to identify key metabolites for growth media design [44] and simulating growth on different substrates with a genome-scale metabolic model to predict media formulation [80]. Efforts that use a metabolic network model must define an *in silico* medium to enable calculations such as Flux Balance Analysis (FBA) [81–83]. The model and simulated medium then are iteratively refined until the network successfully predicts biomass production.

Thus, simulating growth of an uncultured organism with a metabolic model requires the definition of an *in silico* growth medium or a set of candidate media, which may then be validated *in vitro*. The definition of a growth medium *in silico* often begins in the same fashion as *in vitro* attempts: by starting with a medium that has supported simulated growth in models of organisms related to the desired isolate. However, this approach is complicated by the fragmentation of information in the literature. To overcome this obstacle, we have created

MediaDB: a database of experimentally determined, chemically defined growth media conditions that aims to support efforts to leverage -omics data and modeling techniques for characterizing previously uncultured isolates. MediaDB is a manually curated database of defined media formulations for organisms with fully sequenced genomes, emphasizes organisms that have existing metabolic network models, and is the first publically available electronic resource that specifically brings together organisms with genomic data and their associated growth media. MediaDB will facilitate investigation of the relationship between microbial genomes and media composition, serving as both a central repository of data linking genome sequence to media compositions, and as a resource that facilitates model-supported design of cultivation media.

Database construction and content

All data in MediaDB were manually curated from existing primary literature sources. We conducted organism-by-organism literature searches using standard search engines—Google Scholar, PubMed, Web of Science—on the list of *in silico* organisms maintained by the Systems Biology Research Group at UCSD [84]. Our searches were aimed at finding experimentally-verified growth data on defined media for as many organisms with curated metabolic models as possible. The search results were curated manually and the media related information was extracted and formatted in the MediaDB schema, a MySQL database consisting of 12 tables and constructed around 6 main data tables: Organisms, Compounds, Media_Names, Biomass, Sources, and Growth_Data (Figure 2.1). The full schema is included as supporting information (Figure A.1).

Organisms

The Organisms table includes fields for genus, species, and strain, a "type" designation that specifies the organism's kingdom classification, a Boolean value denoting whether the organism has been modeled *in silico*, and, if applicable, a link to the biomass composition for that organism. As shown in Table 2.1, MediaDB currently contains 208 unique Organisms objects spanning 57 species and 46 genera.

Bacteria make up the majority of organisms in the database, reflecting the distribution of species that have been cultured in the laboratory and the MediaDB's emphasis on organisms with existing *in silico* metabolic reconstructions. Such reconstructions exist for 39 of the 43 bacterial species and 51 of the 57 total species in the database. The database also includes many strains for model organisms; *Escherichia coli* and *Bacillus subtilis* contribute 54 and 16 bacterial strains, respectively, to the database.

Compounds

The Compounds table includes fields to describe a chemically-defined compound in terms of its common names, chemical formula, and identifiers that can be used to cross-reference with other databases (KEGG, BiGG, Seed, ChEBI and PubChem)[85–90]. We included identifiers from these databases to enable easier exchange of information between researchers, enhance compatibility with commonly-used resources, and ease development of automated computational analyses that use data in MediaDB. Of the 14,795 compounds contained in the database, 14,785 (99.9%) have identifiers from at least one other database.

Unlike the other tables in the MediaDB schema, the Compounds table was initially curated based on the KEGG database rather than from specific literature sources and was supplemented with

manual entries from other databases as necessary. Its primary purpose is to describe the composition of other data types (Media_Names, Biomass).

Media_Names

The Media_Names table consists of fields specifying a media formulation with a descriptive name, a Boolean value indicating whether or not the particular media formulation was described as minimal in its source material, and a list of names and amounts of each compound that makes up that medium in units of millimolar (mM). Due to the many-to-many nature of relating compounds to different media compositions, the relationship between media formulations and compounds are contained within the Media_Compounds table, but can be queried to find the compounds that make up a particular media formulation. MediaDB only contains chemically defined media formulations and does not include complex formulations, such as media that use yeast extract. The focus on chemically defined media was selected to facilitate computational simulation of growth conditions and to support efforts to cultivate uncultured organisms in the laboratory. MediaDB currently contains 461 different media formulations.

Biomass

The Biomass table consists of fields describing the compounds included in the biomass objective function used in FBA of metabolic network models to simulate exponential cell growth and contains organism genus and species, the list of compounds present in the biomass composition, and the stoichiometric coefficient of each compound in relation to one “unit” of biomass. Like the MediaDB description of media, biomass is also specified by the compounds that make up its composition, resulting in a many-to-many relationship. The Biomass_Compounds table contains

the links between biomass compositions and compounds and can be queried to find the compounds that make up a particular biomass composition.

As detailed in Thiele *et al.* [52], the biomass composition is an important objective function for FBA of metabolic network models; however, it can also be difficult to experimentally determine detailed biomass composition for an organism. Thus, the biomass composition is a salient factor to consider in model construction and refinement, but we found few unique examples of this data type in existing literature sources. Instead, many models have defined the organism biomass composition by using or slightly modifying the biomass objective function from another model. We have included 4 different biomass compositions in MediaDB to provide a basis for users to construct biomass compositions for their own organisms by refining established ones.

Sources

The Sources table consists of fields describing a primary literature source (usually a book or a journal article) and is specified using the first author's last name, the title of the work, the journal, the year of publication and, if applicable, the PubMed identifier and URL to the article. Sources are added to MediaDB if they report experimental laboratory growth of an organism in MediaDB in a medium in MediaDB. MediaDB currently contains 147 unique sources that directly link to any experimental growth media information they provided.

Growth Data

The Growth_Data table describes the combination of physical parameters reported by a literature source for *in vitro* growth of a specific organism. The Growth_Data table links the tables describing an organism, medium, and literature source, and adds information about temperature,

pH, growth rate, product secretion rates, and nutrient uptake rates (whenever reported in the literature source). MediaDB currently contains 765 growth conditions.

In many instances, we found rate data associated with a particular growth condition in the form of an experimentally-measured growth rate (μ) measured in h^{-1}. We stored growth rates in the Growth_Data data field, thereby providing quantitative measures to assist in future metabolic model development. Some growth conditions were also reported with other growth-associated measurements: product secretion rates, medium compound uptake rates and product yields. Unlike growth rates, a growth condition could be associated with multiple measurements of secretion/uptake/yield; hence, we created the Secretion_Uptake table to house these rates and link them to their growth conditions. MediaDB currently contains 557 measured growth rates, 49 metabolite uptake rates, 22 product secretion rates, and 58 product yield coefficients.

Website construction and navigation

The MediaDB website (https://mediadb.systemsbiology.net/) provides a user-friendly interface for performing the two main functions of our database: data browsing and exporting.

Data browsing

Browsing allows the user to query MediaDB with provided data type categories, to manually search through information by navigating through the different data tables or to use keywords to search through the parameters that specify the growth condition entries (see Figure 2.2). The search function matches the given keyword to data entries in all tables and returns the results sorted by the table that contains the matched record.

Tables in the database are linked together on the webpage by cross-referencing to better display all pertinent information for each entry. For example, an entry in the Organisms table shows all of the related growth condition entries collected for that organism, including links to the literature source entries. Similarly, each media formulation entry links to entries for all the compounds present in that media formulation, all of the organisms reported to grow in that media formulation, and the literature source entries where the media formulation was reported. A Compounds entry displays links to all the media formulations in which the compound appears. A Source entry displays links to all the growth conditions reported in that source, as well as links to the online version of that source, when applicable.

Data export

Data can be exported from MediaDB in two different ways, allowing the user flexibility in deciding what information is important for their particular project. The most basic export, found under "Downloads" on the webpage, allows the user to download a copy of the entire MediaDB schema and all database entries to use independently of the website. This option allows the most flexibility in dealing with the data, but requires that the user be familiar enough with relational database management in MySQL to use the SQL file generated by this export.

The second export option is individual media formulation or biomass composition download, available on each media formulation or biomass composition entry page under "Tab-delimited version". This option generates a tab-delimited text file with a list of compounds and their concentrations in the chosen media formulation or biomass composition. The file also includes identifiers for the compounds in other databases. These identifiers facilitate cross-referencing of the various metabolite identifiers used in different *in silico* metabolic network models.

Database utility

Statistics for compounds

Because the MediaDB schema provides links between organisms and the compounds in their growth media, it enables investigation of media components across organisms. For example, we compiled a list of every chemical compound that appears at least once in a growth medium for all 57 species in the database (see Table A.1 for full results). Out of 260 unique compounds, the most commonly occurring compound across all species was calcium chloride ($CaCl_2$), a salt that appears in the growth media of 49 species (86% of all species in MediaDB), because it is often included in stock trace element/mineral solutions. Salts accounted for nine of the top ten most frequent compounds with the only exception being biotin, a vitamin that often appears in stock vitamin solutions and was present for 29 species (51%). Other components of media, such as the carbon source and amino acids, were less uniform across species; the most common carbon source and amino acid were glucose (47%) and cysteine (37%), respectively (a list of the most frequent compounds is shown in Table 2.2).

Our analysis also identified the *least* common compounds in media; 97 of the 260 compounds (37%) appeared in media for only one species and 139 (53%) appeared in media for one or two species only. These uncommon compounds generally fell into one of the following categories: 1) Trace metals included in stock solutions (e.g., nickel sulfate for *Shewanella oneidensis*); 2) Buffers for pH maintenance (e.g., ACES for *Mycobacterium tuberculosis*); 3) Antibiotics used to select for mutant strains (e.g., kanamycin for *Synechocystis PCC6803*); 4) Uncommon carbon sources (e.g., galactose for *Streptomyces coelicolor*); 5) Alternate vitamin forms (e.g., sodium pantothenate rather than calcium pantothenate for *Haemophilus influenzae*); 6) Compounds that

fit niche organism metabolisms (e.g., 2-mercaptoethanesulfonate for *Methanococcus maripaludis*). Compounds in the final category were of particular interest, because they could be tied to unique portions of the known metabolism of the organism. For example, 2-mercaptoethanesulfonate (coenzyme M) only appears in media for the methanogen *M. maripaludis*, because it is a vital cofactor involved in methane production for that organism. As MediaDB grows, we expect that identifying such unusual compounds will play an increasingly useful role in media design.

Linking growth media to metabolism

MediaDB provides a framework for comparing the nutritional requirements of different organisms and currently includes information on a range of microbes, with a focus on organisms that have been modeled *in silico*. In order to demonstrate how MediaDB supports such comparative analysis, we compared media formulations for two organisms that have metabolic network models: *E. coli*, a model bacterium that has been grown with a wide range of compounds (81 different compounds), and *Methanosarcina acetivorans*, a model archaeon that has been grown using a smaller range of compounds (12 different compounds).

Seven compounds appeared in media formulations for both organisms: one carbon source (acetate) and six simple salts (NH_4Cl, $CaCl_2$, $MgCl_2$, KCl, KH_2PO_4, NaCl). The compounds unique to *E. coli* included multiple 5- or 6-carbon sugars (e.g., glucose, lactose, fructose, and succinate) and 19 of the 20 standard L-form amino acids (all except cysteine). The 5 compounds unique to *M. acetivorans* included methanol, a simple carbon source for methanogens that rarely appears in media for other organisms (fellow methanogen *Methanosarcina barkeri* and pathogen *Candida glabrata* are the only other species in MediaDB with media that include methanol). We

also observed that, in contrast to the *E. coli* media data, cysteine was the only amino acid that appeared in growth media for *M. acetivorans.*

We expanded our comparison by using manually curated metabolic models for both *E. coli* [91] and *M. acetivorans* [92] to examine the differences found in media compounds. By examining reactions in the models, we observed that the model for *E. coli* included uptake pathways for many carbon sources that are absent in the *M. acetivorans* model, including all of the carbon sources reported in MediaDB. The *E. coli* model predicted that methanol could be produced during growth, but not consumed, whereas the *M. acetivorans* model predicted the ability to consume methanol for growth and methane production. The models also provided mechanistic justification for our media analysis that suggested differences in cysteine metabolism; the *M. acetivorans* model had the ability to both consume and secrete cysteine and the *E. coli* model predicted cysteine secretion, but not consumption. We extended this analysis by testing the models for growth on a range of experimental media from the database. We selected 11 media for *E. coli*—one for each carbon source—and the one medium for *M. acetivorans* in MediaDB, then simulated each model for growth on all 12 media (see Supplementary File A.1 for an example of this procedure). The *E. coli* model predicted growth on all 12 media, mirroring the organism's versatility to grow on many different carbon sources. The *M. acetivorans* model required modification to remove trace metals from the biomass objective function in order to predict growth on any medium. After the trace metals (which are not included in simulated *E. coli* media) were removed from the *M. acetivorans* model objective function, it accurately predicted growth on its own medium and on the *E. coli* medium with acetate as the carbon source, but not on any of the other media, reflecting the organism's inability to grow on complex carbon sources.

This case study illustrates the use of MediaDB as a tool for investigating the differences in nutritional requirements between organisms and as a source for *in silico* medium formulation. The differences between cultivation media for *E. coli* and *M. acetivorans* were identified using MediaDB and explained using the organisms' respective metabolic models, which include fundamental differences in carbon source and amino acid metabolism. In this example, the results of the comparisons between the media sources and metabolic models were quite parallel, as expected, because both models were manually constructed based on genomic information and information from the primary literature, including media formulation sources. In other cases, where there is disagreement between model simulation results and media information reported, MediaDB will support efforts to improve metabolic network reconstruction by providing information regarding experimentally determined media conditions.

Organism clustering by compound similarity

We used hierarchical clustering of pairwise Euclidean distance between binary vectors of compound inclusion in a medium (e.g., an entry is 1 if a given chemical is included in a medium, or 0 otherwise) to investigate the relationship between organisms in MediaDB based on published growth-supporting media. Figure 2.3 presents a heat map of chemical species in media, created from MediaDB data. The heat map shows bands of high-frequency compounds on the right side of the map and clusters of moderately frequent compounds on the left side; these compound groups are dominated by salts found in stock solutions and L-form amino acids, respectively. The overall sparsity of the heatmap reflects the fact that most compounds occur only once or twice across all species.

We compared this compound similarity tree (Figure 2.3) to a 16s rRNA phylogenetic tree constructed in the Biology Workbench [93–96](Figure 2.4) and found that there was little overlap between genetic similarity and compound similarity. Aside from the two *Methanosarcina* species, which were grown in the same exact media, we observed few parallels between these two trees. Three species in the taxonomic order *Lactobacillales—Lactococcus lactis, Lactobacillus plantarum,* and *Streptococcus thermophilus*—clustered closely together in both trees, but the majority of organisms that formed tight clusters in one tree did not show the same closeness in the other tree. For example, the four *Aspergilli—A. nidulans, A. niger, A. oryzae,* and *A. terreus*—were close in terms of phylogenetic distance, but dissimilar with respect to their media compounds. On the other end of the spectrum, *Corynebacterium glutamicum, A. oryzae, Clostridium beijerinckii,* and *Zymomonas mobilis* show high compound similarity with one another, but are far apart phylogenetically. This observation could be an indication that phylogeny does not correlate to similarity in media formulations, but a more parsimonious explanation is that the data in MediaDB reflect the literature bias towards positive growth results. Due to this lack of negative growth results (i.e. information on what an organism *does not* grow on, which is typically omitted by researchers), we are unable to assert that any organism is incapable of growth in another's media based soley on comparisons of the collected data in MediaDB. This knowledge gap suggests a need for for futher experimental study of the relationship between phylogenetic distance and nutritional requirements for growth. Thus, information available in MediaDB describes whether a given medium has been reported to support a microbe's growth, and may be useful for generating hypotheses of possible media formulations for future experimental efforts. Our analysis also revealed clusters of organisms with high media composition similarity (Figure 2.3) that do not have a clear connection to

observed biology. With further investigation, these similarities could reveal more complex biological relationships that do not fall under the obvious prisms of genetic or environmental similarity. MediaDB will support such comparative studies as the resource continues to grow.

Future development

Community-contributed growth conditions

MediaDB currently contains 57 microbial species, but the scope of the fully-sequenced microbial world is much larger and continues to grow. We intend to expand the breadth of organisms and growth conditions in MediaDB by allowing users to submit their own experimentally verified, defined growth conditions. At this time, we encourage users to submit growth conditions for our review through direct contact with the authors (mediadb@systemsbiology.org), but expect to create an input form that encourages groups to add new data directly through the website.

Analysis tool development

We have demonstrated the potential for media-based comparative analysis using MediaDB with *E. coli* and *M. acetivorans*; however, we have designed MediaDB to support future development of additional tools to support research efforts. We have also made the entire database schema and its contents available for download to further facilitate tool development by MediaDB users. As such tools are developed in our group and others, we will integrate these tools into the website to assist users in their analyses.

Discussion and conclusions

We present MediaDB, a manually curated database of defined media that have been used to cultivate organisms with sequenced genomes. Our database offers several important new

capabilities for researchers through the following features: 1) brings together literature sources of experimentally verified media formulations into a centralized database; 2) contains chemically defined media, so that every compound can be linked to known metabolic pathways in metabolic network models, and so that every formulation is repeatable; 3) links with compound identifiers in existing databases for simple, repeatable and automatable cross-referencing with other sources; 4) focuses on organisms with existing *in silico* models, both encouraging researchers to use and improve such models and providing multiple media conditions to support the iterative development of *in silico* models; 5) serves as a set of organism-specific media conditions to help improve automated metabolic reconstruction methods by replacing more generic media formulations; 6) includes only species with fully-sequenced genomes to ensure that all media formulations can be tied back to genomic data; and 7) is a publically available resource that we expect will grow and increase in usage as growth conditions for more organisms are added. We anticipate that MediaDB will support the investigation of the relationship between organism growth media formulations and genomic information, and facilitate efforts to model microbial metabolism.

Tables and Figures

Kingdom	Genera	Species	Strains
Archaea	*4*	*5*	*14*
Bacteria	*36*	*43*	*172*
Eukaryote	*6*	*9*	*22*
TOTAL	**46**	**57**	**208**

Table 2.1: Taxonomy of organisms currently in MediaDB

Top 10 Compounds		**Top 5 Carbon Sources**	
CaCl2	86%	Glucose	47%
MgSO4	72%	Acetate	33%
KH2PO4	65%	Glycerol	19%
NaCl	65%	Pyruvate	12%
FeSO4	61%	Ethanol/Succinate	9%
ZnSO4	61%	**Top 5 Amino Acids**	
K2HPO4	58%	Cysteine	37%
NH4Cl	53%	Aspartate	33%
Biotin	51%	Arginine	33%
CuSO4	49%	Glutamate	32%
		Leucine	32%

Table 2.2: Highest frequency compounds in MediaDB. Percentages reflect the fraction of species that contain each compound in at least one growth medium.

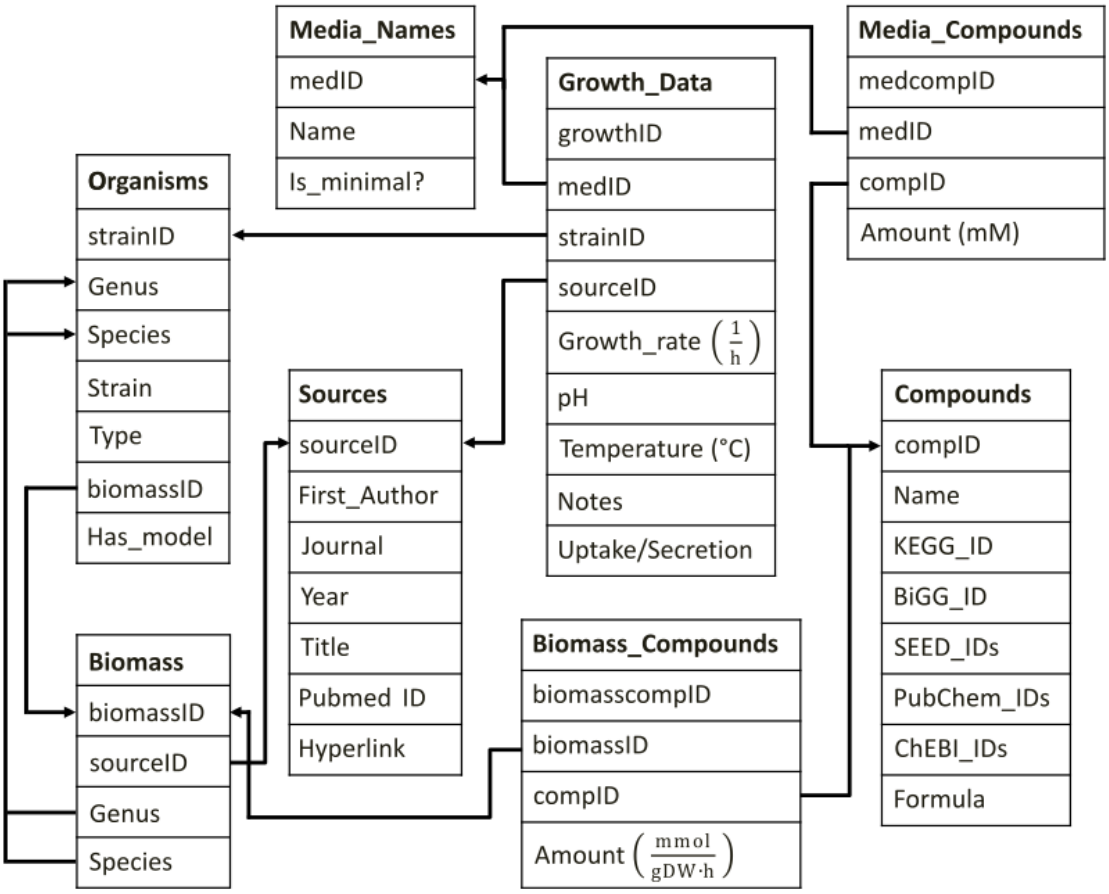

Figure 2.1: Simplified database schema. This graph shows the connections between the 6 main tables, Organisms, Compounds, Media_Names, Biomass, Sources, and Growth_Data. Also shown are Media_Compounds and Biomass_Compounds, linking tables that connect the Compounds table to the Media_Names and Biomass tables, respectively. Arrows indicate foreign key relationships, in which the head of the arrow points to the primary key being referenced. A full map of the MediaDB schema containing all tables and their connections can be found in Figure A.1.

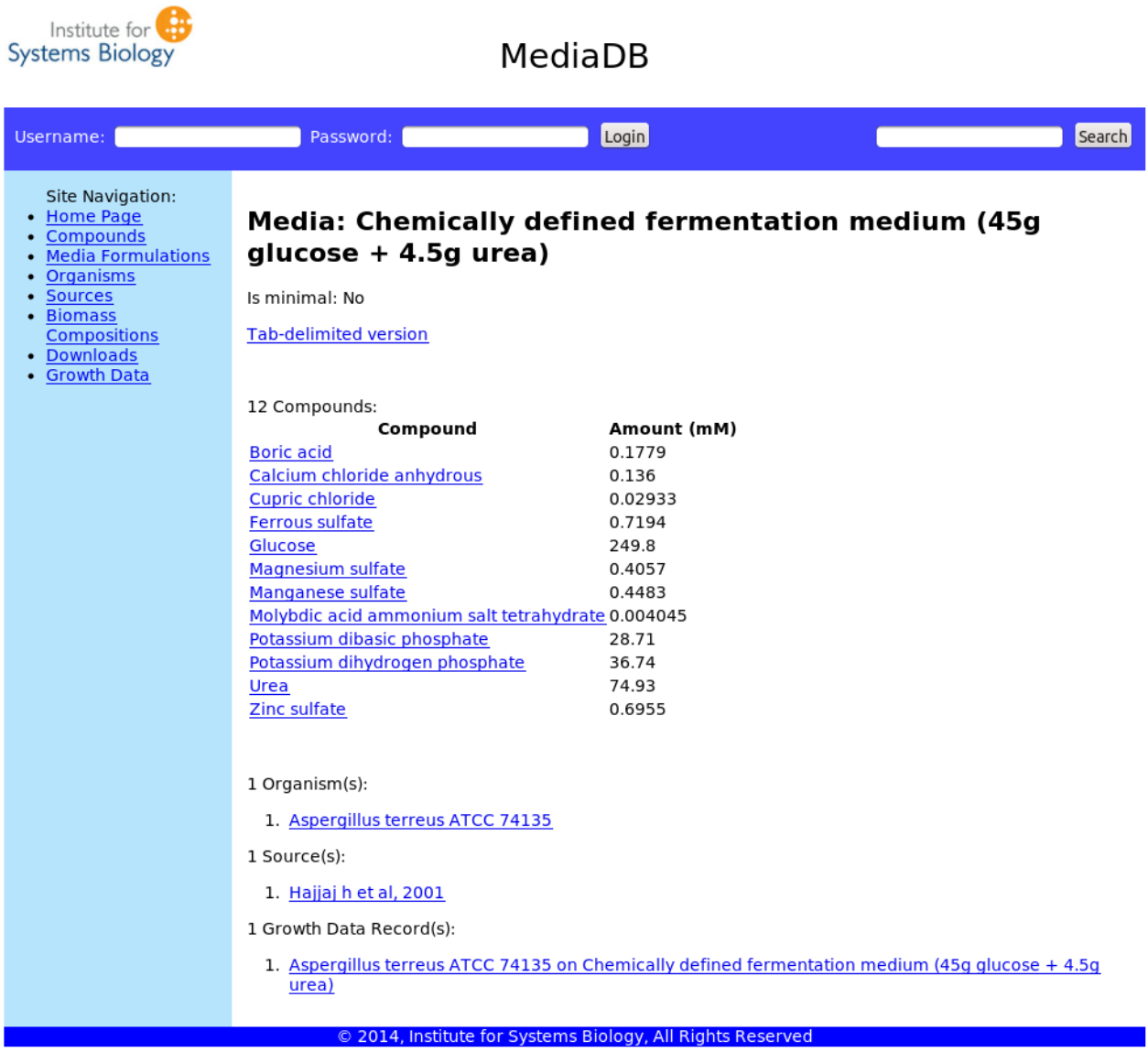

Figure 2.2: The MediaDB website. The database can be found at https://mediadb.systemsbiology.net. This page shows the composition of a media formulation and displays links to the organism, source, and growth record that use this medium. The "Site Navigation" panel lists the different tables that can be browsed manually and also the "Downloads" tab, where the user can export a copy of the entire MediaDB schema. The search field is at the top right of the page.

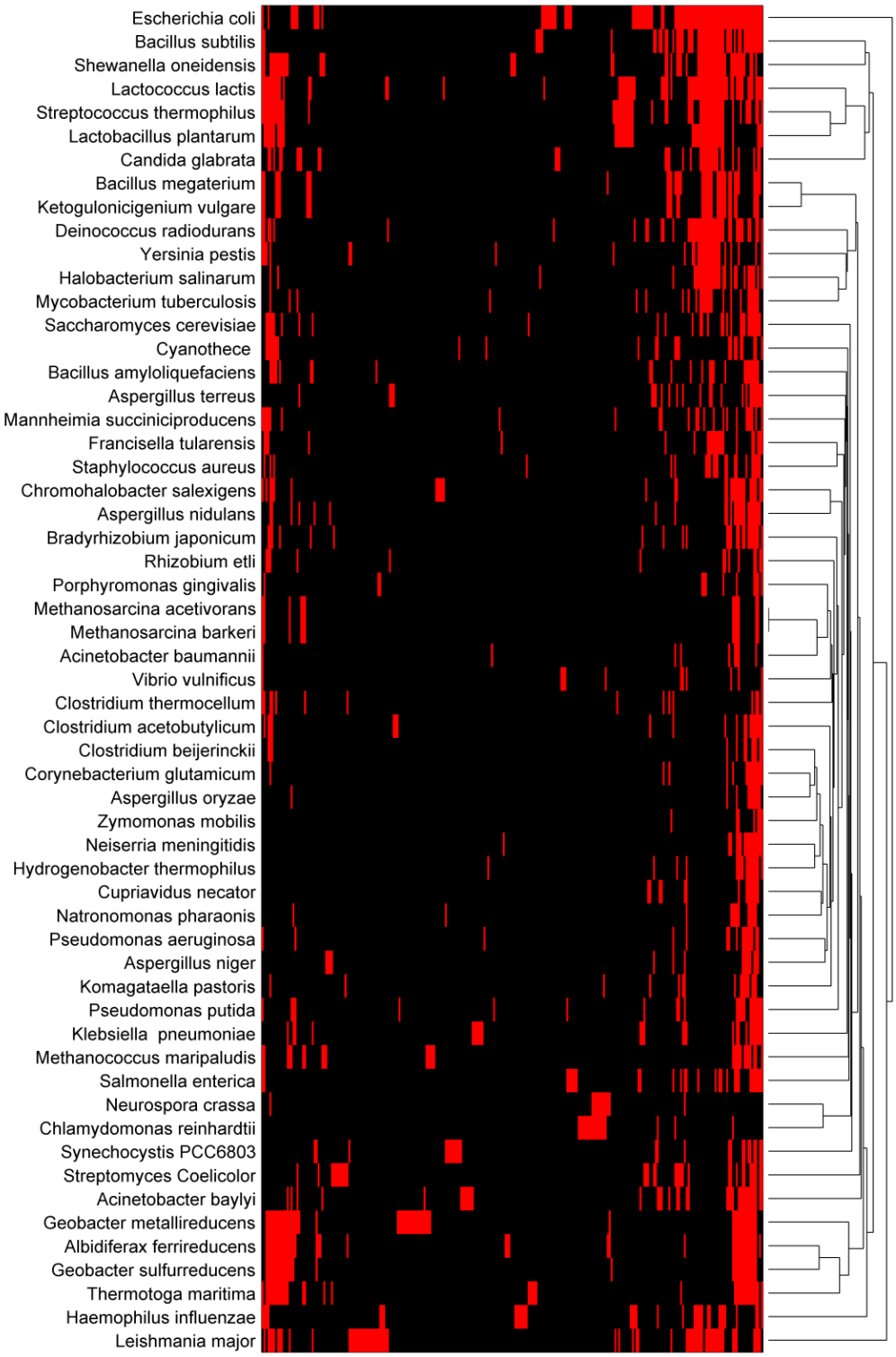

Figure 2.3: Heat map and dendrogram showing hierarchical clustering of species based on media compositions. Red bars indicate compounds that occur in at least one medium for that species. Black bars indicate compounds that do not appear in any media for that species. This figure was generated using the Statistics Toolbox in Matlab.

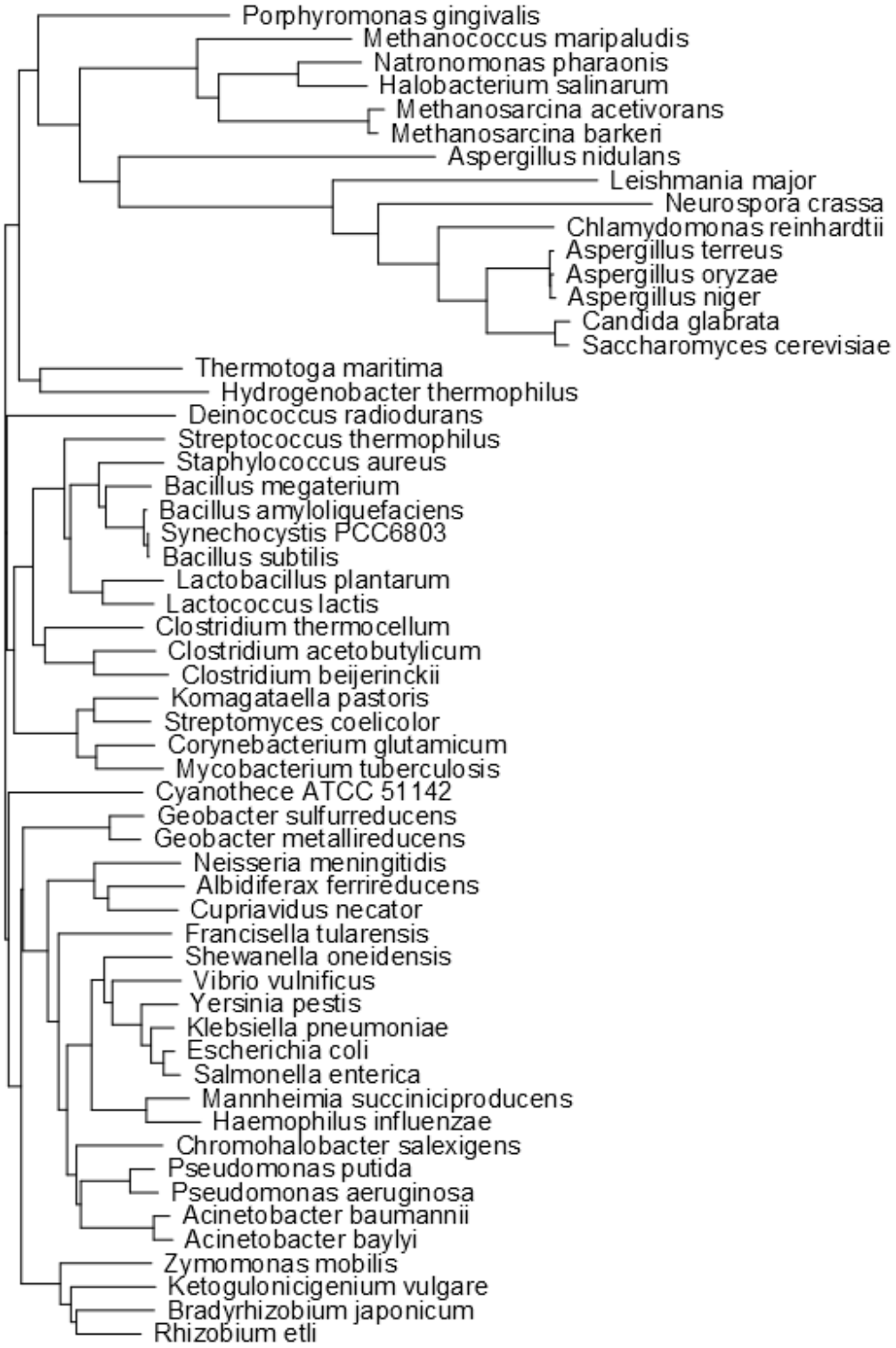

Figure 2.4: Phylogenetic tree of 16S rRNA sequences for species in MediaDB. Phylogeny was inferred from a CLUSTAL W alignment generated in the Biology Workbench using 16S rRNA sequences from the SILVA database.

Chapter 3: Exploring Hydrogenotrophic Methanogenesis: A Genome Scale Metabolic Reconstruction of *Methanococcus maripaludis S2*[2]

Introduction

Methane plays a critical role in the global carbon cycle and as a greenhouse gas, is 21 times more potent than carbon dioxide [97] in absorbing and emitting energy. Additionally, it is a candidate bridge fuel [98] because it burns comparatively cleaner than traditional fossil fuels. Advancing technology also enables this gas to be converted to high energy density liquid fuels with a lower carbon footprint [99]. Methane is the second most abundant greenhouse gas after carbon dioxide [100] and is produced in the environment by biological and non-biological sources [101]. Methanogens are the largest biological contributors of methane, producing about 1 Gt of methane gas per year [102]. This group of microorganisms from the domain Archaea grow on carbon dioxide or one or two carbon compounds using enzymes containing unique biological co-factors [103,104].

Though phylogenetically and metabolically diverse, methanogens can be separated into two groups based on the presence or absence of cytochromes [102]. The cytochrome-lacking methanogens (sometimes referred to as hydrogenotrophic methanogens) mainly use H_2, and sometimes formate, as sources of electrons for CO_2 reduction to methane. In contrast, cytochrome-containing (or methylotrophic) methanogens utilize acetate and methylated compounds for methanogenic growth with a minority also being able to use H_2 and CO_2. Although both groups have similar central pathways

[2] This chapter contains material from an article in preparation for submission to *Journal of Bacteriology.* I would like to thanks my coauthors: Thomas Lie, Juan Zhang, John Leigh, and Nathan Price

of CO_2 reduction, they possess differing modes of energy coupling [105] at the last methanogenic step involving heterodisulfide reductase (Hdr).

The reduction of the CoM-S-S-CoB heterodisulfide with H_2 or reduced electron carriers is exergonic and can be directly or indirectly coupled to energy generation. In the methylotrophic methanogens, a membrane-associated cytochrome-containing Hdr (HdrDE) receives reducing equivalents from a methanogen-specific membrane-soluble electron shuttle, methanophenazine, for reduction of the heterodisulfide. This results in proton extrusion and the creation of a membrane potential for ATP generation [106,107]. However, in the hydrogenotrophic methanogens, the Hdr (HdrABC) is cytoplasmic and no membrane potential is generated. Instead, Hdr mediates a bifurcation of electron flow in which the exergonic heterodisulfide reduction is coupled to and drives the endergonic reduction of a ferredoxin used for the first step of methanogenesis [108].

Methanococcus maripaludis [109] belongs to this group of hydrogenotrophic methanogens. Compared to the larger genomes of methylotrophic methanogens, its genome is relatively small and contains only 1722 protein coding genes [110]. It grows robustly with a doubling time of 2 hours [109] and is genetically tractable [111], and thus has been an ideal candidate for studying methanogenesis, unique co-factors and their biosyntheses [112], and gene regulation [113].To avoid environmental fluctuations that can affect gene regulation, a system for continuous culture of *M. maripaludis* [114] has been established for steady state transcriptomic [115] and proteomic [116] studies of *M. maripaludis* strains. Several groups have also employed larger systems biology approaches to perform predictive studies using this organism [117]. With these tools in place, and the ability for expression of heterologous genes in *M. maripaludis* [118,119], the metabolic engineering of *M. maripaludis* for various industrial use is the obvious next step.

Genome scale metabolic reconstructions are powerful tools that map and elucidate metabolic pathways. They are organism-specific knowledge bases that can be used for simulating steady state growth via flux balance analysis (FBA) [83] by generating constraint-based models. Using these models, we can hypothesize different metabolic scenarios that can then be tested experimentally. They have helped guide metabolic engineering efforts to produce industrial biochemicals in multiple organisms [120,121]. Similarly, a genome scale metabolic reconstruction for *M. maripaludis* would not only promote a better understanding of methanogenesis but also support metabolic engineering efforts that could harness the unique metabolism of this hydrogenotrophic methanogen. Other groups have already created metabolic models of *M. maripaludis*; as part of a mutualistic community model with *D. vulgaris* [122] and under axenic conditions [123]. In the former case, the model of *M. maripaludis* represented only core metabolism and was used primarily to investigate interactions between the two different species rather than probe the depths of the organism's metabolism [122]. The latter case was the first genome-scale metabolic reconstruction of *M. maripaludis* [123], an important step towards understanding *M. maripaludis* metabolism.

In our model, iMR540, we made important updates and refinements to various pathways based on recent literature. The most critical was the electron bifurcation step that has been described above as it explains the ability for this organism to grow despite the lack of a proton-exporting electron transport chain. This also includes eliminating methanophenazine utilization and synthesis, which is part of the membrane bound electron transport system of the methylotrophic methanogens and is absent in hydrogenotrophic methanogens [102]. Additional features include a corrected sulfur assimilation pathway [124], and the addition of various biosynthesis pathways for all of the unique coenzymes involved in methanogenesis [125]. We increased genome coverage by employing likelihood-based gap filling, a technique that fills reaction gaps based on gene homology rather than on parsimony [126]. Our

reconstruction is the first manually-curated genome scale reconstruction to employ likelihood based gap filling. Furthermore, we expanded the scope of our reconstruction beyond stoichiometric considerations by creating a new method to approximate overall model free energy. This is an especially salient consideration for methanogenic archaea, which can grow close to the thermodynamic limits that support life [127]. A well-established method of applying free energy constraints involves applying the second law of thermodynamics to metabolic models to restrict reaction directionalities in the direction of negative free energy generation [128,129]. Rather than apply thermodynamic constraints to every metabolic reaction, we created a method that predicts overall free energy generated during steady state growth based solely on standard free energies and effective concentrations of external metabolites. In combining these novel thermodynamic considerations with stoichiometric information, iMR540 provides a means to predict energetically feasible strain designs, enhancing our metabolic engineering capabilities with *M. maripaludis.*

Methods

Genome Scale Reconstruction Procedure

The process of genome scale metabolic network reconstruction has been reviewed previously [52] and begins with annotating an organism genome using gene-protein-reaction (GPR) relationships stored in a reaction database. Several databases are available for this purpose [10,51,130]; we chose the Department of Energy Systems Biology Knowledgebase (Kbase; www.kbase.us), a suite of tools that includes the Model SEED reaction database [51]. We created our first draft reconstruction using the stored Kbase genome for *M. maripaludis S2* (genome id: kb|g.575) and the automated reconstruction method ("Reconstruct Genome-scale Metabolic Model"). For this initial reconstruction, we used the default gram negative biomass composition and filled knowledge gaps using likelihood based gap filling

(method currently not supported in Kbase Narrative Interface). This yielded the first full draft of the metabolic reconstruction that could predict growth when simulated as a model.

We expanded and refined the model by manually adding information from literature sources. In cases where reactions from literature were part of the Model SEED database, we labeled the reactions using SEED identifiers, names, subsystems, and EC numbers. For other cases where we encountered reactions that were not part of the Model SEED we created unique reaction identifiers and names, then added subsystem information based on our knowledge of the metabolic network. We also adhered to SEED identifiers, names, formulas, and charges for metabolites whenever possible and had very few cases where we specified our own values. Metabolites were compartmentalized using standard tags for cytosol ("c0") and extracellular ("e0") compartments. These tags additionally identify *M. maripaludis* as "Organism 0" in the possible future case where we could add other organisms to create a community metabolic reconstruction. Exchange reactions used for introducing metabolites to the extracellular compartment were standardized in "EX_{metabolite ID}[e0]" format. Comprehensive information on the reactions, metabolites, and genes in our reconstruction can be found in Supplementary Materials.

Model Simulations with Flux Balance Analysis

To make rigorous quantitative growth predictions, a genome scale metabolic reconstruction can be simulated as a model. Reactions and their participating metabolites in the metabolic network are connected via the stoichiometric matrix (S), which contains the stoichiometric coefficients for each metabolite (row) in each reaction (column). The S-matrix is used as the basis of a model via the principles of metabolite mass conservation by recognizing that time-dependent accumulation of metabolites in the system (b) is equivalent to the product of the S-matrix and the vector of reaction fluxes (v)

$$Sv = b \quad [1]$$

In flux balance analysis (FBA), we further simplify this differential system by assuming our organism is in steady state growth; thus b=0 and the system is linear [81]. This assumption bounds our model system to a large solution space that can further be constrained by applying upper and lower bounds to each reaction flux:

$$v_{i,lower} \leq v_i \leq v_{i,upper} \quad [2]$$

To find feasible flux distributions that represent likely physiological states within this solution space, we solved our model by optimizing the biomass objective function, a simulation of maximum cell growth yield [131]. We further constrained possible flux distributions by minimizing the squared sum of fluxes, effectively forcing our model to find solutions that minimize the total flux in the system while maximizing growth. All model simulations were performed using the COBRA toolbox 2.0 [132] in MATLAB [7.14.0.739] (The MathWorks Inc., Natick, MA).

To encourage model transparency [133] and assist future users in simulating condition-specific models, we created several functions that create these models, simulate maximum growth with the aforementioned constraints, and print relevant information from the flux distribution (Supplementary File B.1). We also wrote numerous functions to help modify the reaction network, retrieve specific useful pieces of information from model simulations, and diagnose issues that may arise during model use. For several of these functions, we used the Paint4Net toolbox [134] to draw flux maps that show the direction and magnitude of fluxes in a given FBA solution. A limited number of our functions are included with this manuscript in their current versions (Supplementary File B.1) with the full set of up-to-date tools available on Github (https://github.com/marichards/methanococcus).

Gene Knockout Phenotype Simulations

Because a model is based around the stoichiometry of reactions contained in the S-matrix, knocking out a gene is akin to knocking out all reactions that depend on the gene. Thus, performing a gene knockout phenotype simulation in a metabolic model requires that model reactions be linked to genes via GPR relationships. We performed gene knockout simulations using our function "simulateKOPanel.m" (Supplementary File B.1), which relies heavily on the "deleteModelGenes.m" function in the COBRA Toolbox 2.0 [132] as well as several of our own functions. Our experimental test set included 18 knockout genotypes across 4 different growth conditions, with 30 total wet lab experiments across these conditions [135–140]. We simulated growth phenotypes for all 72 combinations of knockout genotypes and growth conditions and then evaluated these growth phenotypes as lethal/non-lethal with a threshold of 10% wild type growth. Predictive accuracy was assessed by comparing predictions on the 30 known phenotypes with wet lab data. We further evaluated our model's performance using the Matthews correlation coefficient (MCC), a metric that evaluates correlation based on a -1 to 1 scale [141]:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad [3]$$

Thermodynamic Calculations

We added standard free energies of formation (1 mM, 25ºC, 1 bar, pH=7, ionic strength = 0.1 M) from the Equilibrator database [142] to all exchange reactions for which these values could be reliably estimated via the group contribution method [143]. To incorporate these values into our reconstruction, we expanded the standard model structure to include a "freeEnergy" numerical array with length equal that of the "reactions" array. For calculating overall free energy of a flux distribution, we created an "optimizeThermoModel.m" code (Supplementary File B.1) that is built around the

"optimizeCbModel.m" code in the COBRA Toolbox 2.0 [132]. Our script accepts effective concentrations (mM) for specified exchange metabolites, assumes standard activities of 1 mM for unspecified metabolites, and uses these values to calculate effective metabolite free energies based on the reconstruction's stored values for each exchange reaction. Prior to performing FBA, we add these free energies to the exchange reactions, which ordinarily have the form:

$$A_{e0} \rightleftharpoons$$

We alter these exchanges such that production of a metabolite "creates" free energy equivalent to the metabolite's free energy of formation:

$$A_{e0} \rightleftharpoons \left(\Delta G_{A_{e0}}\right) dG$$

Here, $\Delta G_{A_{e0}}$ is the stoichiometric coefficient of a new metabolite "dG" that is used to sum model free energy. Because exchange reactions must satisfy mass balance by necessarily entering or exiting the model without creating new metabolites, adding free energies to the model creates an imbalance that we must correct. We restore model balance by allowing "dG" to exit the model via its own exchange reaction (GIBBS_kJ_GDW):

$$dG \rightleftharpoons$$

Measuring the total flux of the exchange reaction gives an estimation of total free energy being generated in an FBA solution on a per cell mass basis. We have incorporated this thermodynamic calculation into all of our available model simulations (Supplementary File B.1); thus by default, we calculate and print overall model free energy in every flux distribution.

Dry Cell Weight and Growth Yield Measurements

Wild type *M. maripaludis S2* cells were grown in McNA medium—a chemically defined medium for growth on H_2 and CO_2 supplemented with acetate (Table B.3)—using a 1-L chemostat under anaerobic conditions as described previously [114]. The chemostat was operated in steady state continuous mode under H_2-limiting conditions to match model simulation conditions, with gas flows of 10-20 mL/min H_2, 40 mL/min CO_2, 15 mL/min of a H_2S:Ar mixture (1:99 v/v), and a balance of N_2 up to a total 200 mL/min. We altered our growth rate of *M. maripaludis* during steady state by varying pump speeds to achieve dilution rates of approximately 0.045-0.090 h-1, checking OD660 periodically to ensure steady state at each data point. For each sample point, we measured growth rate based on dilution rate and methane evolution rate via a combination of a bubble flow meter to assess total gas outflow and a Buck Scientific model 910 gas chromatograph equipped with a flame ionization detector to quantify methane fraction.

We recalculated calibration curves for dry cell weight versus optical density by measuring dry cell weight via cell filtering and OD_{660} via a UV/Vis spectrophotometer {Spectronic 20D+} blanked with water. After measuring chemostat optical density, we sampled 50 mL aliquots of cells in suspension directly from chemostat culture and centrifuged samples at 7000 RPM for 15 minutes. 40 mL of supernatant was removed by pipette, then cells were re-suspended in the remaining 10 mL of media. These concentrated aliquots were vacuum filtered through 0.45 µM pore filters to remove all non-cellular components, then dried at room temperature and weighed daily until their weights stabilized.

Growth yields were calculated based on doubling time (t_d, equal to ln(2) x (dilution rate x 60)$^{-1}$) as described previously [144], but with our measured conversion between OD_{660} and cell density:

$$Y_{GDW/CH4} = \frac{OD_{660}}{CH_4(\frac{mL}{min})} \times \frac{0.46 g/L}{1\ OD_{660}} \times \frac{1}{t_d(min)} \times \frac{22{,}400\ mL}{mol}$$

ATP Maintenance and Predicted Growth Yields

As described by Thiele and Palsson, the optimal way to obtain accurate ATP maintenance values is to plot ATP production versus growth data from chemostat growth experiments [52]. In practice, this requires measuring steady state growth rate in concert with an uptake rate or, in our case, a product secretion rate, as described above.

To calculate ATP maintenance values in our model, we constrained our model to our measured growth rate and methane secretion rate at each sampling point and set the model objective to maximize ATP hydrolysis (rxn00062[c0]). We plotted each resulting value of ATP production as a function of growth rate and obtained the growth-associated (slope) and non-growth associated (y-intercept) ATP maintenance values using a linear model, as described by Thiele and Palsson [52]. The resulting plot can be found in Figure B.4.

Our growth data points comprised a set of 9 measurements and we used them as both training and test data by employing leave one out cross validation (LOOCV). In the LOOCV approach, a set of N samples is divided into a training dataset of N-1 points and a test sample of 1 point. The model developed on the training set is then tested on the remaining point that was left out of the training data. In employing this method, we iteratively removed one point from our full dataset and determined ATP maintenance values for that N-1 dataset as described above to create a trained model. We then constrained our model's methane secretion flux to the measured rate in the remaining test point and predicted maximum growth rate within that constraint using our trained model. Using these values, we calculated predicted growth yields for each point using the above formula and compared them to our measured values for each point. All simulations were performed using the default H_2 + CO_2 media formulation supplemented with acetate (McNA medium).

Reconstruction and Model Availability

Reconstructing a metabolic network is an iterative process and therefore, it is paramount that reconstructions be as clear as possible to encourage future updates and expansions [133]. We have strived for clarity in both our nomenclature and in our decision making process for including each reaction present in our reconstruction. Reactions and metabolites in our network are based upon identifiers and names found in Kbase, but also include crosslinks to ChEBI [145] and KEGG identifiers [10], enzyme commission numbers, and reaction subsystems where available. Each reaction in the reconstruction is also connected to its literature and/or database source, plus its reaction confidence score when applicable (Table B.1).

Additionally, we have sought to maximize usability of both our reconstruction and our model. The systems biology markup language (SBML) is a standard medium for distributing metabolic reconstructions [34]; thus, we have included our reaction network in SBML level 2, the highest version currently supported by the COBRA Toolbox [132]. In our experience using reconstructions from other groups, we have found a wide range of usability, from those that can easily be imported and simulated to those that are difficult to use and interpret. In the interest of making our simulations and results easy to reproduce, we have included our reconstruction in MATLAB data structure format and an example of our codes for simulating model growth on different media and gene knockout phenotypes (Supplementary File B.1).We have also made our codes and reconstruction available on Github (https://github.com/marichards/methanococcus).

Results

Reconstruction Statistics

The basic statistics for iMR540 are displayed in Table 3.1. Notably, reactions are categorized as 1) internal reactions, occurring entirely within the cytoplasm; 2) transport reactions, involving

translocation of at least one chemical species across the cell membrane; 3) exchange reactions, which supply metabolites to or remove metabolites from the model. Of the 586 internal reactions in our network, over 85% of the internal reactions are associated with at least one gene. We suspect that a major reason for this high percentage of gene-associated reactions was our use of likelihood based gap filling, which resulted in the automated addition of 66 genes to our reconstruction before manual curation. Furthermore, we relied heavily on biochemical knowledge from literature sources, particularly regarding recently-elucidated biosynthesis pathways that were not initially available in annotation databases. Our combined use of maximum likelihood gap filling and reliance on published literature sources are the likely causes for our consistent ties to gene homology.

Another salient detail of our reconstruction is that it includes many "dead-end" metabolites and reactions that cannot be synthesized or consumed. Thus, these metabolites and reactions are not part of our simulatable model, but we have included them in our reconstruction because they are all gene-associated; all dead end internal reactions in our reconstruction have at least one gene association. This indicates that there is genetic evidence supporting the presence of each dead end reaction and metabolite, thus they should be involved in metabolism even though we have not yet elucidated full synthesis or consumption pathways. They represent excellent candidates for further exploration of *M. maripaludis* metabolism, particularly as iMR540 is updated and expanded in the future.

Conversely, our reconstruction contains 86 internal reactions that lack genes, many of which were added during automated gap filling but some of which were added manually. All of our reactions are annotated with subsystems, allowing us to assess where each reaction fits into metabolism, including those without genes. Figure 3.1 shows the breakdown of reactions without genes, where the subsystems have been manually grouped into broader categories (e.g. "Amino Acid Biosynthesis" instead of "Glycine Biosynthesis"). The largest group of these reactions is the "Unique Coenzyme

Syntheses", which includes reactions that synthesize coenzyme M, coenzyme B, tetrahydromethanopterin (H_4MPT), methanofuran, coenzyme F_{420}, and coenzyme F_{430}. Although these 24 reactions lack genes, all of them were added manually as hypothetical steps to complete essential biosynthetic pathways and are based on information from biochemical literature. These are distinct from, for example, the 11 reactions encompassed by "Vitamin and Cofactor Synthesis" that were added to fill biosynthesis gaps but have no supporting literature evidence. We expect that as experimental research groups uncover more biochemical phenomena, they will determine genes that tie to the reactions in the former group. The gap filling reactions, much like dead end reactions and metabolites, point us toward poorly-understood areas of metabolism in our organism and require more investigation into both the reaction pathways and their associated genes.

As an additional feature of our reconstruction, our use of likelihood based gap filling also assigned likelihood scores for many of the reactions in the reconstruction. These confidence scores quantify the probability of a given reaction being part of the metabolic reconstruction on a scale of 0-1 and provide a new metric of evaluating our confidence in the reconstruction. We can then use the scores to quickly hone in on both reactions that lack genes and gene-associated reactions with low gene homology as possible targets for more experimental investigation. They also provide a logical starting point for future users looking to expand upon and improve the existing reconstruction.

Electron Bifurcation and Acetate Metabolism

Methanogenesis from H_2 and CO_2 has often been represented as a linear pathway with heterodisulfide reduction as the final step. This was demonstrated to be mediated by methanophenazine dependent membrane bound heterodisulfide (HdrDE) [106,107] for the cytochrome containing methanogens. However, the non-cytochrome containing obligate hydrogenotrophs do not contain the typical membrane associated heterodisulfide reductase but instead one that is most likely associated with the

cytoplasm [146,147]. Additionally, it is a three subunit complex (HdrABC) with a flavin adenine dinucleotide (FAD) containing co-factor in the HdrA subunit [148].HdrA and other FAD-containing enzymes have been increasingly recognized as sites for electron bifurcation, coupling an exergonic reaction with an endergonic reaction in a two-step transfer of one electron [149,150]. Recently, it has been demonstrated [108,151] that this heterodisulfide reductase mediates the coupling of exergonic heterodisulfide reduction with endergonic ferredoxin reduction. As shown in Figure 3.2, this ferredoxin is used for reduction of the CO_2 via Fwd thereby linking the last step of methanogenesis with the first step in a cyclical fashion [152]

The assumption of a linear pathway in *M. maripaludis* without accounting for electron bifurcation can affect the downstream predictions in the metabolic model. The default mechanism of energy conservation in our initial, uncurated model matched methylotrophic methanogens and utilized methanophenazine, an electron carrier known to be absent from *M. maripaludis* and other hydrogenotrophic methanogens. We removed methanophenazine-based electron flow and added the correct electron bifurcation pathway, linking heterodisulfide reduction with electrons from H_2 to carbon dioxide reduction via reduced ferredoxin instead. This commonly-encountered reconstruction pitfall, in which information available in annotation databases does not sufficiently represent recently elucidated metabolic pathways, emphasizes the need to keep abreast of updated academic literature in spite of the improvement of automatic reconstruction methods.

To demonstrate that the linear pathway cannot support growth of *M. maripaludis*, we altered the native electron bifurcating heterodisulfide reductase (HdrABC) reaction:

$$CoB - S - S - CoM + 2\,H_2 + Fd_{ox} \rightleftharpoons HS - CoB + HS - CoM + 2\,H^+ + Fd_{rd}$$

by removing ferredoxin, balancing mass and charge to yield:

$$CoB - S - S - CoM + H_2 \rightleftharpoons HS - CoB + HS - CoM$$

This scenario represented a hypothetical case where *M. maripaludis* does not contain a membrane-bound HdrDE complex but cannot perform electron bifurcation. We optimized this altered model for growth on CO_2 + H_2 and were unable to predict *in silico* growth, supporting the observation that the ferredoxin reduction via electron bifurcation is an essential part of our network. Lack of model growth can clearly be attributed to disruption of the central energy coupling mechanism in *M. maripaludis*, in which electron bifurcation must necessarily reduce ferredoxin for reducing CO_2. The alternative source of reduced ferredoxin is the energy-converting Eha hydrogenase, which utilizes a sodium ion gradient to reduce ferredoxin with H_2 on a 1:1 basis. CO_2 reduction to methane requires reduced ferredoxin and pumps out sodium ions, also on a 1:1 basis. Thus, each cycle of methanogenesis in this scenario effectively produces no sodium ion gradient for synthesizing ATP, the central component necessary for biomass formation. Additionally, methanogenesis loses small amounts of carbon for biosynthesis; hence, reducing one ferredoxin effectively pumps less than one sodium ion across the cell membrane and creates an overall energy deficit. Overall, this simulation illustrates the essentiality of ferredoxin reduction via electron bifurcation and reinforces the idea that Eha hydrogenase can play only an anaplerotic role in methanogenesis [136].

Taking this analysis one step further, we used our reconstruction to probe into acetate assimilation, a pathway in *M. maripaludis* that can enhance growth but cannot replace H_2 and CO_2 as an energy source [153]. This is in contrast to multiple methylotrophic methanogens such as *Methanosarcina barkeri* that can subsist using solely the aceticlastic pathway [154]. It is unknown why *M. maripaludis* cannot be grown on acetate alone, and our reconstruction did not reveal any strictly stoichiometric obstacle to growth. However, much like the pathway in *M. barkeri,* an aceticlastic pathway in *M. maripaludis* would require energy-converting hydrogenases (Eha and Ehb) to produce H_2 using reduced ferredoxin,

pumping out sodium ions, and thrusting this reaction into a central stoichiometric role rather than an anaplerotic one. As shown in Figure 3.3, when we simulated our model and allowed Eha/Ehb unlimited flux, we could predict aceticlastic growth with Eha/Ehb oxidizing approximately two moles of ferredoxin per methane produced. We then constrained our model to enforce a solely anaplerotic or biosynthetic role of energy-converting hydrogenase by limiting flux through the Eha/Ehb reaction to 10% that of methane secretion rate. Doing so prevented our model from predicting growth from acetate alone, but did not restrict hydrogenotrophic growth or supplementary acetate uptake. This simulation supports the hypothesis that *M. maripaludis* cannot achieve aceticlastic growth because Eha or Ehb cannot assume a central role in methanogenesis. In keeping with this hypothesis, we have restricted flux through Eha/Ehb in our model to ≤ 10% of methane secretion as a default constraint.

Interestingly, there is evidence that *M. maripaludis* uses multiple forms of ferredoxin as electron carriers and may link certain steps, particularly those involved in electron bifurcation, reduction of CO_2 to formylmethanofuran, and certain biosynthetic reactions, using specific ferredoxins [155]. Presently, the full extent of this phenomenon is not well understood and requires more experimental investigation. However, in an effort to represent ferredoxin specificity in our model, we have included a function (see https://github.com/marichards/methanococcus) that replaces promiscuous ferredoxins with two types of specific ferredoxins. One type is used for the Eha hydrogenase, Hdr, and formylmethanfuran dehydrogenase (Fwd) and the other type for Ehb hydrogenase and biosynthetic carboxylating oxidoreductases, as suggested by [135]. Using this function tightens the coupling between the aforementioned reactions by restricting each set to one pool of electron carriers and allows us to predict how ferredoxin specificity could change possible model flux distributions. In wild type simulations, this change has minimal effects on predicted growth yields and fluxes but could have notable impact on gene knockout predictions, particularly those involving reactions that utilize ferredoxin. Moreover,

electron movement through different ferredoxin species could have important implications for hypothesizing strain designs, thus including multiple ferredoxins could be vital for effective metabolic engineering.

Other Biochemistry Improvements

A major part of our manual curation was adding biosynthesis pathways for the methanogenic coenzymes, sugars, and lipids. *M. maripaludis* utilizes a number of unusual coenzymes (methanofuran, H_4MPT, coenzyme F_{420}, coenzyme B, coenzyme M, coenzyme F_{430}) as carbon and electron carriers during methanogenesis [156]. It also contains recently characterized pathways for synthesizing a tetrasaccharide for N-linked glycosylation of archaellin (archeal flagellin) [157] and multiple forms of archaeol, an archaeal membrane ether lipid [158]. None of these pathways were included in our draft reconstruction and few were completely present in the Model SEED database, thus the bulk of these reactions were added manually. These synthesis pathways, particularly for coenzymes, are vital pieces of *M. maripaludis* metabolism that set it apart from the vast majority of known biochemistry. Hence, including these synthesis pathways and adding these metabolites as required biomass components was crucial for distinguishing our reconstruction from existing networks.

In a similar vein, we sought to accurately represent sulfur assimilation, a pathway not yet fully understood in *M. maripaludis*. Sulfate is known not to be the sulfur source for *M. maripaludis*; moreover, sulfate reduction would produce sulfite, a methanogenesis inhibitor [159]. However, because sulfate is the default sulfur source for most microorganisms, our first draft reconstruction included a sulfate transporter and sulfate reduction pathway. We removed this default pathway and instead added a pathway to utilize hydrogen sulfide gas, the primary sulfur source for *M. maripaludis.* Our updated sulfur assimilation pathway includes sulfide oxidation to sulfite—an essential metabolite for multiple biosynthetic pathways—via a hypothesized dissimilatory sulfite reductase-like protein [124]. Taken

together with aforementioned syntheses, these modifications demonstrated the need for rigorous manual curation to add known biochemical pathways that were not part of the automated reconstruction and remove pathways that are known not to function in the organism. By employing these methods and by working collaboratively with *M. maripaludis* experts, we have created a reconstruction that maximizes consistency with biochemical literature of our organism.

Growth Yield Validation and ATP Maintenance

Evaluating a metabolic network reconstruction by qualitatively comparing it to known biochemical phenomena is a valuable way to gauge how close the network can represent actual biochemistry. To make more quantitative comparisons, we must convert the reconstruction to a metabolic model by imposing flux constraints on the network, enforcing mass balance on all metabolites, and optimizing to an objective function (Methods). A common way of quantitatively evaluating the resulting model is to simulate maximum cell growth under steady-state conditions and compare growth yield predictions to experimentally-determined values. Due to the narrow range of possible substrates for our hydrogenotrophic system and scarcity of growth yield data for our organism, we generated our own experimental measurements of growth yield. We conducted chemostat growth experiments under H_2-limiting conditions and measured growth yields as described previously [144], but varied our dilution rate to gather a range of different yield measurements. Cell density was assessed using optical density (OD) and was previously reported as OD_{600}=1 corresponding to 0.34 mg(dry weight)·ml^{-1} [137]. We were unsure of the efficacy of this value, in part because we measured at 660 nm rather than 600 nm. We re-measured this conversion factor using a combination of centrifugation and vacuum filtering (Methods) and plotted a new calibration curve (Figure B.3), determining that OD_{660}=1 corresponded to 0.462 ± 0.015 mg(dry weight)·ml^{-1}. Using this value, we calculated measured growth yields based on growth

rates (equal to dilution rates) and measured methane evolution rates (Methods). Measured growth yields are plotted in Figure 3.4 for 9 independent steady state time points.

We then tested our model by generating growth yield predictions and comparing them to measured growth yields. Growth yield predictions depend not only on metabolic steps where ATP is generated or hydrolyzed, but more heavily on ATP maintenance energies [160]. From a modeling perspective, maintenance energies are regarded as the moles of ATP needed to support cellular processes not otherwise depicted in metabolism, including DNA replication, RNA transcription, protein synthesis, and other requirements. We recognized that our model was essentially untrained in terms of ATP maintenance and contained automated values from our first draft reconstruction. Thus it was crucial to train our model by fitting to our experimental dataset. However, we were also wary of overfitting our model by training and testing on the same set of samples. We addressed both concerns by performing leave one out cross validation (LOOCV) on our full dataset. Thus, for each of our nine growth rate values, we used the remaining eight growth rates and their associated measured methane evolution rates to derive ATP maintenance values. We then used that ATP maintenance value in our calculation of predicted growth yield for the given growth rate. Using this method allowed us to essentially test our model's growth yield predictions on each separate test point while training on the remaining 8 measurements. The resulting predicted growth yields are plotted in Figure 3.4 along with our measured growth yields. As illustrated by this plot, our model was able to consistently predict growth yield within the 95% confidence interval of a measured test sample after being trained on a separate dataset. Though growth yield validation is not an absolute measure of model performance, our model's ability to closely reproduce experimental results in a LOOCV setting that mitigated overfitting suggested a high propensity for generating viable growth predictions. Moreover, the relative consistency between

measured and predicted values indicated our model's robustness for predicting growth yields across a range of different dilution and methane secretion rates.

We also used the full dataset of growth rates and methane evolution rates to set final values for growth associated maintenance (GAM) and non-growth associated maintenance (NGAM). The GAM represents ATP hydrolysis required to support growth-related processes and NGAM represents ATP hydrolysis required for non-growth associated cellular upkeep. GAM was originally set as 40.11 (mmol per grams [cell mass]), a relatively low value when compared with that of a fast-growing bacterial species, such as the GAM of 59.81 in *E. coli* [161]. NGAM, represented by simple ATP hydrolysis, was unbounded in our first draft reconstruction and took on a value of 0 during all model simulations. After training on our full dataset, we set our GAM and NGAM values to 169.9 mmol ATP per gram [cell mass] and 5.0 mmol ATP per gram [cell mass] h^{-1}, respectively (see Appendix B). Notably, these maintenance values are much higher than those in other methanogen models; for example, fellow methanogen *Methanosarcina barkeri* was reported to have a GAM of 65.00 (mmol per grams [cell mass]) [160], about 38% of our calculated value. This difference is reflective of the observed differences in growth yield for these organisms during growth on H_2 and CO_2. Using the same formula for growth yield in each case at nearly identical doubling times of 12 h, *M. maripaludis* grew at a yield of about 33% of that reported for *M. barkeri* [160]. Thus, though we calculated unusually high ATP maintenance requirements for growth, these high values reflect observed differences in growth data when comparing to a methylotrophic methanogen growing on the same substrates.

Gene Knockout Validation

Gene knockout experiments present a different method for validating a metabolic reconstruction based on its model. At its core, a constraint-based model is built around gene-protein-reaction relationships that connect genotype to growth phenotype. Thus, comparing model predictions of gene knockout

lethality provides an excellent way to quantitatively measure the qualitative content of the model. This process hinges on the availability of gene knockout data for the organism being modeled, ideally with the abundance of data found for a traditional model organism such as *Escherichia coli* [162]. *M. maripaludis* lacks this abundance of *in vivo* gene knockout data, but was used for transposon mutagenesis to calculate an essentiality index of all of its genes [163]. Although this dataset does not contain the same quality of knockout data as actual knockout experiments, it provides a valuable "first pass" test set for gene essentiality of our model. However, essentiality index is itself a model for predicting gene knockout lethality, thus although we compared our model's predictions to this dataset (see Appendix B) it did not provide the same clear picture as targeted knockout experiments.

Because much of methanogenesis revolves around the function of different hydrogenases, the bulk of available gene knockout data involves hydrogenase knockouts on different media. For our test set, we were able to assemble a knockout panel of 30 binary growth phenotypes based on previous publications [135–140]. Though the breadth of these knockout genotypes is limited, they are all vital pieces of central carbon metabolism and therefore, they give us a good idea of how well our model can predict knockouts in central catabolism. In comparing with these data, as shown in Figure 3.5, our model achieved 90% prediction accuracy and a Matthew's correlation coefficient of 0.67. These high values suggested that our model is an excellent predictor of growth phenotype based on genotype changes in central carbon metabolism. This result was particularly encouraging because we avoided training our model on this dataset in the interest of preventing overfitting our model to the validation set.

It is also worth noting that all 3 incorrect predictions have similar bases in the model. In these cases, knockouts of 5 or 6 hydrogenases are experimentally found to be lethal in formate-grown cells, or in formate + CO-grown cells lacking carbon monoxide dehydrogenase (CODH), yet our model predicts these knockouts to be non-lethal. The reason for this disagreement lies in our innate assumption that

every reaction performs at 100% efficiency, an ideal scenario that is not achievable in an actual organism. Methanogensis cannot be expected to operate at 100% enzyme efficiency, as some of substrates and electron carriers will not react, thus it can be considered as a "leaky" process where a portion of the metabolites are unused in every cycle. Specifically, in the $\Delta 5H_2$ase and $\Delta 6H_2$ase knockouts, small amounts of hydrogen are synthesized in biosynthetic reactions. Eha hydrogenase remains active in each mutant and can use this hydrogen to supply anaplerotic reduced ferredoxin for methanogenesis. However, in reality an additional non-stoichiometric amount of hydrogen is required. Thus, the actual mutants cannot grow on formate alone and require hydrogen. [Notably, most of our knockout predictions were made with glyceraldehyde-3-phosphate ferredoxin oxidoreductase (GAPOR) constrained to carry zero flux. The GAPOR reaction is ferredoxin-reducing and can serve as a supplemental source of reduced ferredoxin for growth on formate in the case of Eha knockout [138]. However, in wild type strains the expression of GAPOR is not sufficient to support growth in the absence of other hydrogenases (e.g. the $\Delta 5H_2$ase and $\Delta 6H_2$ase mutants). As demonstrated previously, overexpression of the GAPOR operon allows for growth of these mutants ($\Delta 6H_2ase_{supp}$ and $\Delta 7H_2ase_{sup}$) on formate [138]. To best reflect these genotypic differences, we altered the bounds of the GAPOR reaction (rxn07191[c0]) in our knockout simulation code, constraining the reaction to zero flux in all cases except those of the $\Delta 6H_2ase_{supp}$ and $\Delta 7H_2ase_{sup}$ mutants.]

Thermodynamic Calculations

Free energy plays a key role in biochemistry as all biological systems must have a sufficiently low overall free energy to support growth. When simulating optimal growth using a metabolic model we expect the same rules to apply to our system, hence we can apply thermodynamic constraints to the model based on metabolite free energies of formation. In a previous study, free energies of formation were used to constrain reversibility of all internal model reactions based on the second law of thermodynamics [128].

This method, while rigorous, is highly dependent on concentration and can be overly restrictive with regard to predicted flux distributions; thus it is most effective when paired with metabolite effective concentration data [129]. Lacking extensive effective concentration data for *M. maripaludis,* we chose to represent free energy constraints in a novel approach where we add free energies only to exchange reactions, the set of metabolites that can be taken up or produced by the model. These metabolites effectively represent the organism's overall biochemical "reaction"; therefore it is reasonable to expect this overall reaction must produce a negative overall free energy to support growth. Indeed, applying this method to our default model growing on H_2 + CO_2 with methane evolution rate of 50 mmol/g(dry weight)·h, overall free energy production is predicted as -5.59 kJ/g(dry weight). Optionally, this calculation can be used as an additional model constraint that restricts overall free energy to be negative, the equivalent of imposing the second law of thermodynamics on the organism itself.

We expect that this straightforward calculation (Methods) will be a useful addition to our model, particularly as we aim to use it as a platform for generating possible strain designs. With regard to free energy, methanogens are particularly notable in that they subsist close to the thermodynamic limit to support growth [127]. It follows that for any potential strain design, we must pay particular attention to the overall free energy of our system, lest it dip below this vital threshold. It may also provide a metric for differentiating between multiple feasible strain designs by ranking them in order of thermodynamic feasibility. At the very least, it serves as an additional capability of our model and as a checkpoint to ensure that our overall stoichiometry matches up with overall free energy. We have included example functions for adding metabolite free energies to our model and performing FBA with an additional free energy calculation (Appendix B).

Discussion

Genome scale metabolic reconstructions provide a wide lens for studying the biochemical complexity in a computational setting. We used likelihood based gapfilling and meticulous manual curation to build iMR540, a comprehensive reconstruction of *M. maripaludis* that incorporates electron bifurcation to portray cyclical hydrogenotrophic methanogenesis. We incorporated many unique pathways that differentiate our network from those for other organisms, creating a novel tool for understanding and probing more deeply into hydrogenotrophic methanogenesis. The resulting network model compared favorably with measured growth yield and gene knockout data and provided a platform to develop a new method for estimating overall free energy generation during steady state growth.

Electron bifurcation is the central energy conservation mechanism in *M. maripaludis,* thus it is fitting that this process takes a central role in our reaction network. This mechanism is in stark contrast to existing methanogen models that contain linear methanogenesis based on oxidative (electron transport) phosphorylation [92,123,160]. While the linear model is correct for methanogens with cytochromes, it is not correct for methanogens without cytochromes such as *M. maripaludis*. We have demonstrated that, in the absence of a membrane-bound HdrDE complex, ferredoxin reduction via electron bifurcation is essential for predicting growth in our network. Furthermore, constraining the energy-conserving Eha/Ehb reaction to a minor metabolic role provides a stoichiometric hypothesis for the inability of *M. maripaludis* to grow aceticlastically and will undoubtedly influence model predictions moving forward. Ferredoxin specificity for these and other reactions remains an open question that could profoundly affect electron carrier utilization and have implications in native and mutant genotypes, a possibility we have acknowledged by allowing either promiscuous or specific ferredoxins in our reconstruction.

Beyond bifurcation itself, we added numerous uncommon biosynthetic pathways to our network from literature sources that further separate it from models of other organisms. These pathways included

syntheses for methanogenic coenzymes, archaellin sugars and archaeol lipids as well as a relatively novel sulfur assimilation pathway. Additionally, using likelihood-based gap filling helped us automatically identify 66 more genes, increasing the gene coverage of our reconstruction prior to the start of manual curation and assigning reaction likelihood scores for many reactions that lend a measure of confidence level to network. The efficacy of these methods is shown not only in the qualitative accuracy of our reconstruction, but also in the formidable quantitative capabilities of the resulting model. Our model performed well in a LOOCV analysis of growth yield data and compared favorably with experimental gene knockout data, suggesting a high propensity for generating predictions that are consistent with observed biology.

For a methanogen living close to the edge of thermodynamic feasibility, we also thought it salient to include some calculation of overall free energy when simulating our model. We have thus introduced a novel method of predicting overall model free energy generation based solely on standard free energies and concentrations of exchange metabolites. Though a relatively trivial calculation, our method gives a quick assessment of whether a predicted flux distribution is thermodynamically possible and could prove a particularly useful tool for guiding future metabolic engineering designs.

While considering our reconstruction's consistency with existing literature and our model's high performance on measured data, it is poignant that we acknowledge the limitations in our network. First, though we have attempted to address as many parts of metabolism as possible, many "dark areas" of *M. maripaludis* metabolism still exist in our reconstruction. For many of these cases, gene annotations from Kbase and likelihood based gap filling give us starting hypotheses for what may be occurring in these dark areas, but the accuracy of these predictions remains unknown until they have been biochemically characterized. We recognize that our reconstruction effort represents only an incremental step toward understanding *M. maripaludis* metabolism and that many other users may follow in our

footsteps. With these considerations in mind, we strived for maximum transparency in our metabolic network to make our reconstruction decisions apparent to future users and to make our results easily reproducible. There is ample opportunity for improving our reconstruction in the future by elucidating the missing information for these dark areas and we hope that by providing information on the origins and likelihoods of our reactions, we can encourage exploration of these as-yet-unknown pathways.

Second, we recognize that even for the areas of metabolism that we understand well, our model is purely stoichiometric and therefore can only provide predictions from a metabolic perspective. This somewhat limits the scope of questions we can ask using our reconstruction because it does not explicitly include information for other cellular processes, e.g. transcriptional regulation. Given the wide expanse of unknown metabolism, we do not perceive this limitation as particularly crippling, as we can still ask a plentiful supply of questions just within the realm of stoichiometry. In the future, if we wish to address this limitation our stoichiometric predictions could be combined with those from other types of structures, thus providing the tools to probe questions that include other cellular processes.

Lastly, we stress that even within the metabolic space, our model's power lies in predicting the scope of metabolic possibility, not absolute biological reality. Any particular flux distribution should be considered a hypothesis about the what our organism can theoretically achieve, not a precise prediction about all metabolic fluxes. These predictions provide valuable insight into the potential metabolic capabilities of our organism, but it would be folly to accept any single prediction as a facsimile of reality. Such a consideration is vital when considering our model or any other model as a tool for facilitating metabolic engineering designs because any model prediction should be considered as a starting point rather than a final product. By explicitly acknowledging this limitation, we hope to realistically portray the capabilities of our reconstruction as a tool to better understand the unique biochemistry of

hydrogenotrophic methanogens, push forward biochemical discovery in these organisms, and unlock their potential as metabolic engineering targets.

Tables and Figures

Methanococcus maripaludis S2 model statistics	
Protein Coding Genes	540
% ORF Coverage	31
Intra/Extracellular Metabolites	658/53
Dead End Metabolites	259
Internal Reactions	586
Transport/Exchange Reactions	49/59
Dead End Reactions	206
Gene-Associated Reactions	500

Table 3.1. General statistics for the iMR540 reconstruction.

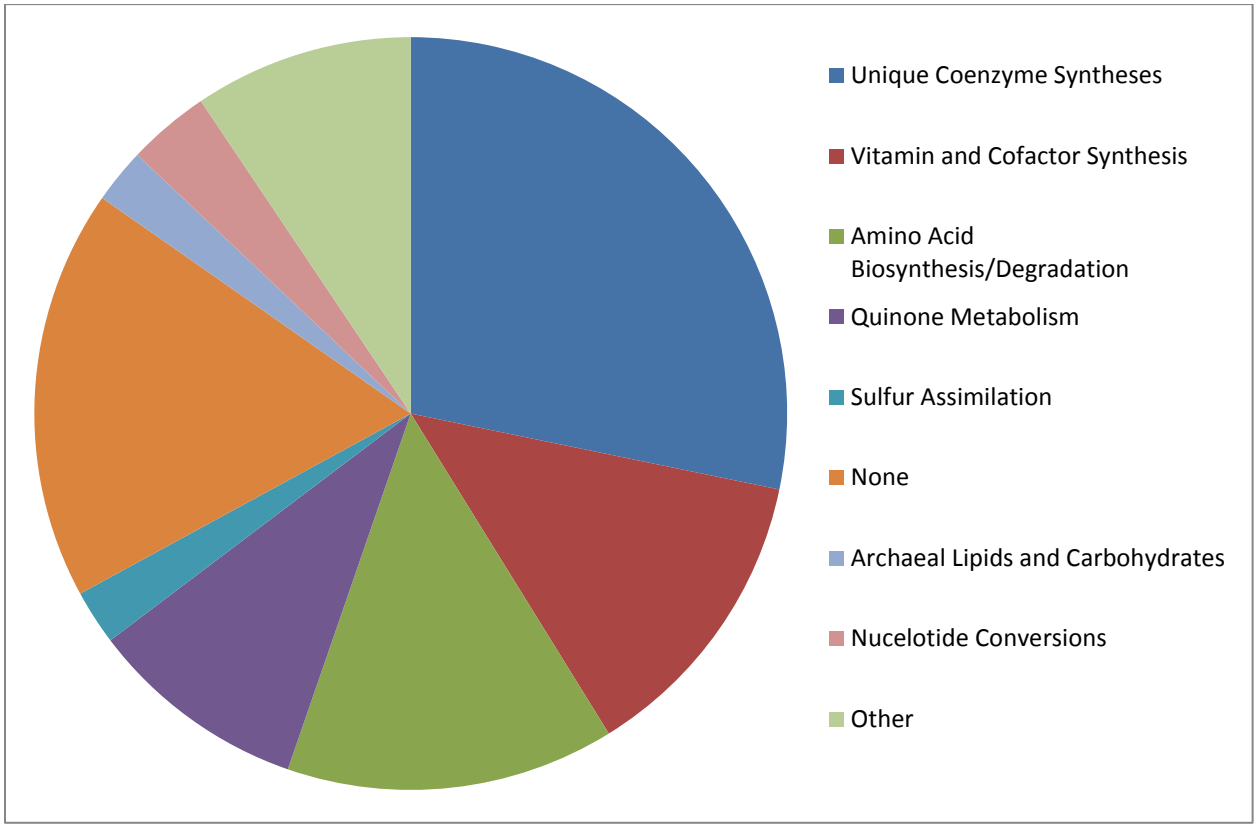

Figure 3.1: A chart showing broad subsystem groupings of the 85 reactions in iMR540 that are not associated with any genes. Reactions falling underneath the "None" subsystem grouping were present in the Model SEED database but had no subsystems listed there and no obvious membership in another subsystem. Reactions grouped within "Other" were dissimilar both from the other categories and from one another, thus we felt they did not merit creation of multiple additional categories.

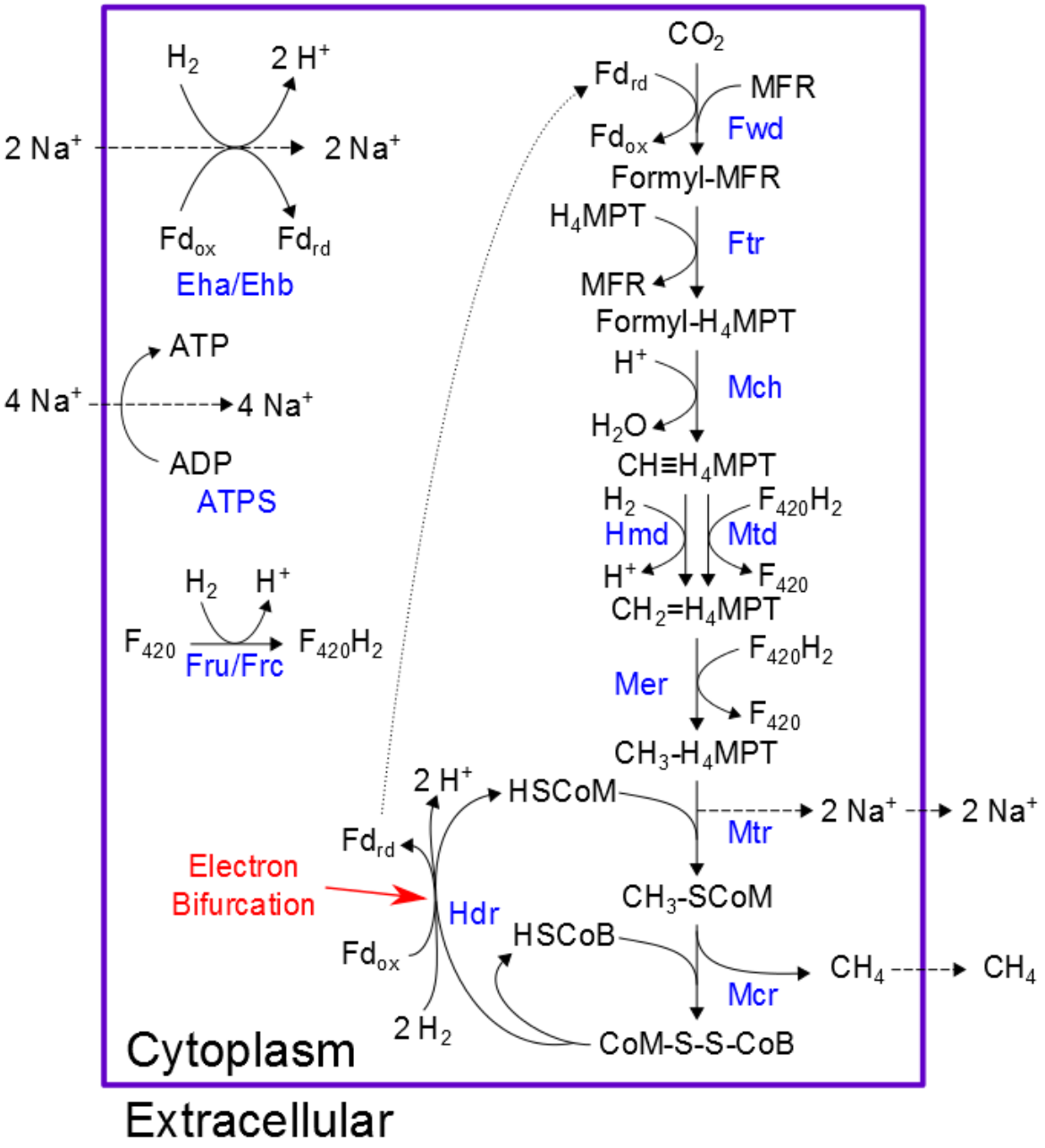

Figure 3.2: The native pathway of hydrogenotrophic methanogenesis present in *M. maripaludis.* As shown, electrons from 2 moles of H_2 are split between reducing ferredoxin and regenerating coenzymes B and M. Reduced ferredoxin from this reaction links it to CO_2 reduction, the first step in the pathway. Enzyme names are shown in **blue**. Metabolites: Fd_{rd}, reduced ferredoxin; Fd_{ox}, oxidized ferredoxin ; MFR, methanofuran; HSCoM, coenzyme M; HSCoB, coenzyme B; F_{420}, coenzyme F420. Enzymes: Fwd, formylmethanofuran dehydrogenase; Ftr, formylmethanofuran/H_4MPT formyl transferase; Mch, methenyl-H_4MPT cyclohydrolase; Hmd, H_2-dependent methylene-H_4MPT dehydrogenase; Mtd, F_{420}-dependent methylene-H_4MPT dehydrogenase; Mer, methylene-H_4MPT reductase; Mtr, methyl-H_4MPT coenzyme M methyltransferase; Mcr, methyl coenzyme M reductase; Hdr, heterodisulfide reductase;Eha/Ehb, energy-conserving hydrogenases; ATPS, ATP-synthase; Fru, F_{420}-reducing hydrogenase (selenocysteine-containing); Frc. F_{420}-reducing hydrogenase (cysteine-containing).

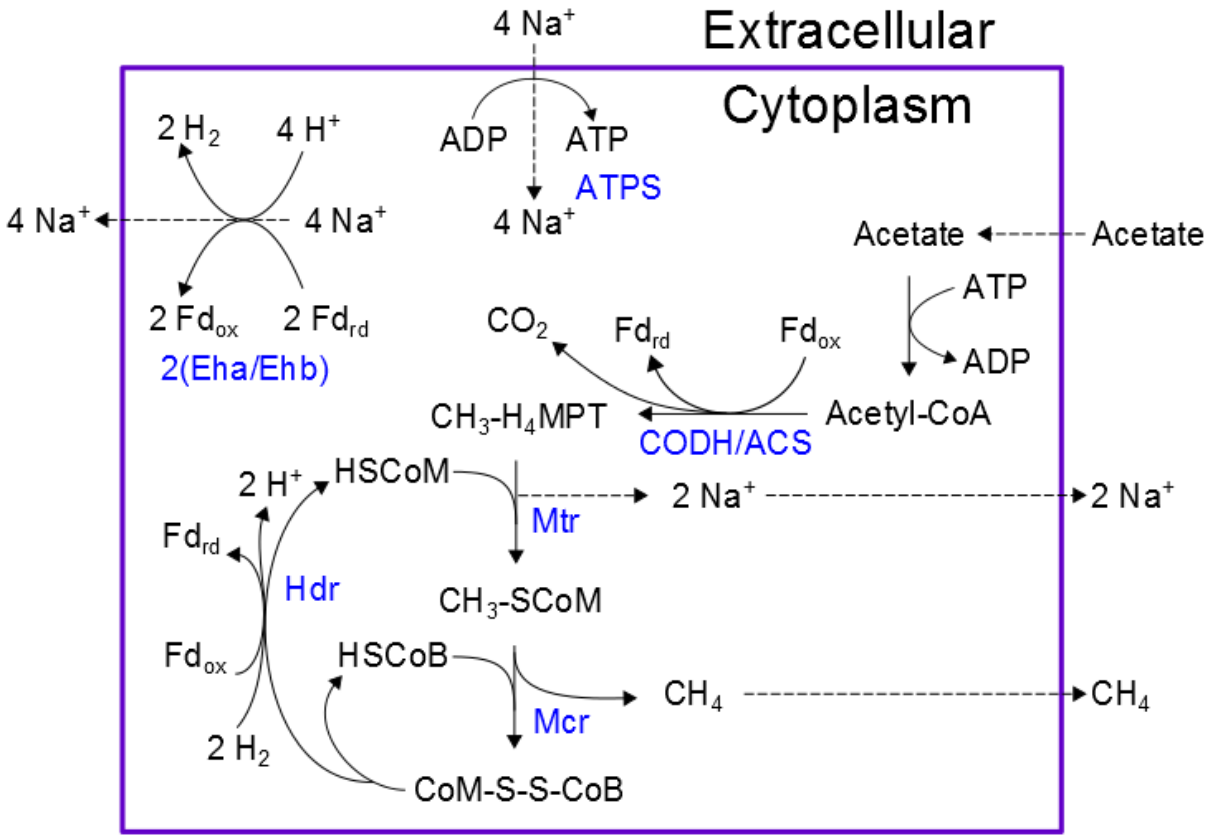

Figure 3.3: Hypothetical pathway for aceticlastic methanogenesis in *M. maripaludis*. As demonstrated, this scheme would require 2 cycles of Eha/Ehb in order to oxidize ferredoxin reduced by the CODH/ACS and Hdr reactions. By constraining the Eha/Ehb reaction to only 10% of methane efflux, this pathway becomes infeasible. Enzyme names are shown in **blue.** Metabolites: Fd_{rd}, reduced ferredoxin; Fd_{ox}, oxidized ferredoxin ; MFR, methanofuran; HSCoM, coenzyme M; HSCoB, coenzyme B; F_{420}, coenzyme F420. Enzymes: CODH/ACS, carbon monoxide dehydrogenase/acetyl-CoA synthase complex; Mtr, methyl-H_4MPT coenzyme M methyltransferase; Mcr, methyl coenzyme M reductase; Hdr, heterodisulfide reductase;Eha/Ehb, energy-conserving hydrogenases; ATPS, ATP-synthase.

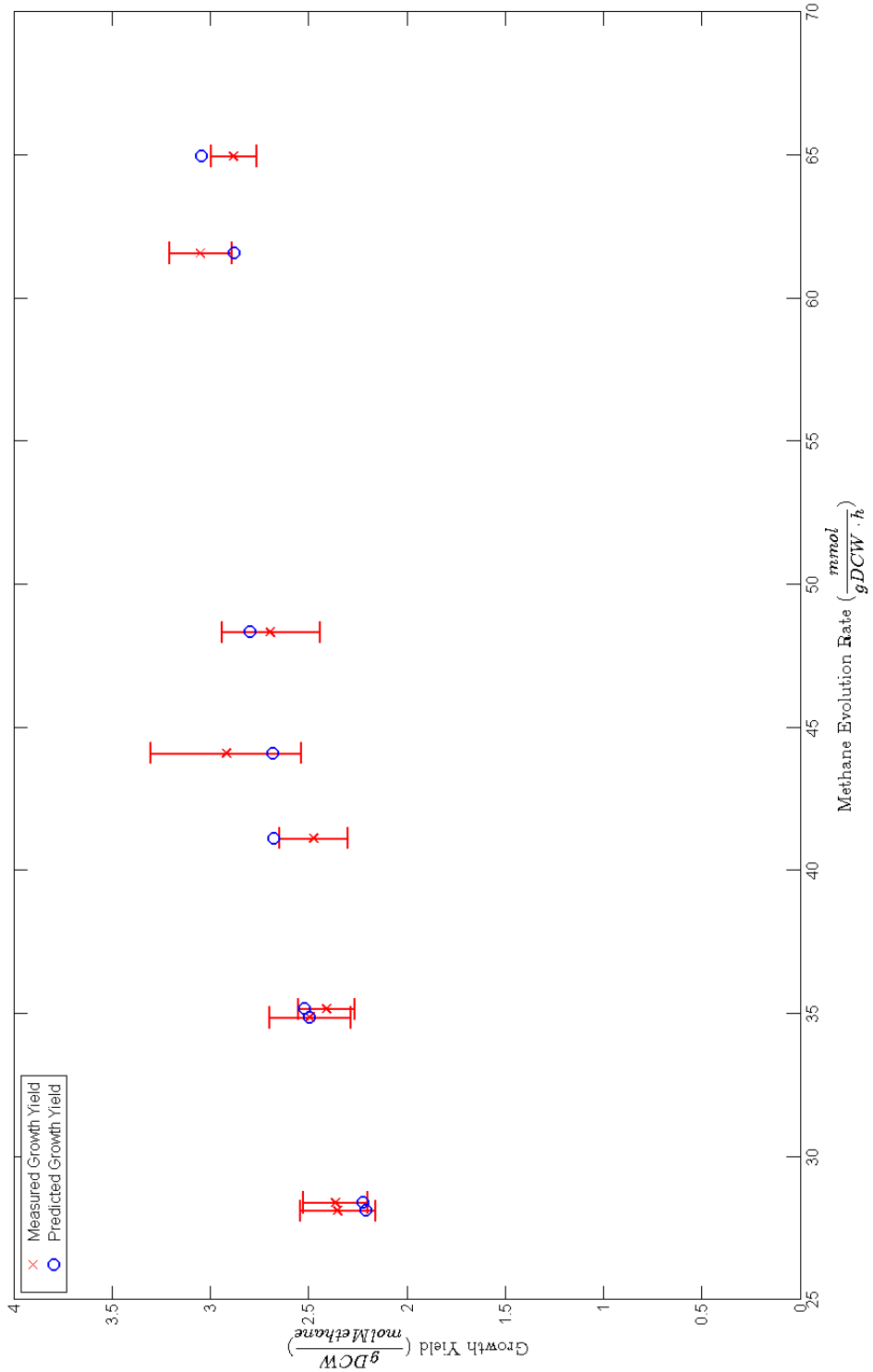

Figure 3.4: Comparing growth yield predictions on hydrogen to measured data using LOOCV (Methods). All but two predicted growth rates fall within the 95% confidence interval of the measured values. Each of the two outlying points are predicted to grow to higher than measured growth yields.

Genotype	H_2	Formate	H_2 + Formate	Formate + CO
Δhmd	N	N	N	N
Δmtd	N	N	N	N
ΔfrcA	N	N	N	N
ΔfruA	N	N	N	N
ΔfrcAΔfruA	N	N	N	N
ΔvhcAUΔvhuA	N	N	N	N
ΔhdrB2	N	N	N	N
ΔfdhA1	N	N	N	N
ΔfdhA2	N	N	N	N
ΔfdhA1ΔfdhA2	N	L	N	L
ΔfdhA2ΔfdhB2	N	N	N	N
ΔehbF	N	N	N	N
Δ3H2ase	N	N	N	N
Δ5H2ase	L	N	N	N
Δ6H2ase	L	N	N	N
Δ6H2aseΔcdh	L	N	N	N
$\Delta 6H2ase_{supp}$	L	N	N	N
$\Delta 7H2ase_{supp}$	L	N	N	N
Total Correct:	**10 of 10**	**14 of 16**	**2 of 2**	**1 of 2**

Figure 3.5: Knockout lethality predictions from running FBA on our models show close agreement with experimental results of hydrogenase knockouts. Green boxes indicate growth phenotypes where our models correctly replicated experimental results; red boxes indicate growth phenotypes where our models were incorrect; white boxes indicate growth phenotypes where we lacked experimental validation data. Across the full spectrum of conditions, our models correctly predicted 27 of 30 conditions (90%) accurately, resulting in a strong Matthews Correlation Coefficient of 0.67. This suggests that our reconstruction produces models that accurately depict the effects of genotype alterations on growth phenotypes. L = lethal, N = non-lethal.

Chapter 4: Guiding Strain Design with the iMR540 Metabolic Reconstruction

Introduction

There is currently great interest in harnessing non-sugar feedstocks to produce liquid fuels and value-added chemicals. One such example is converting methane—a greenhouse gas—to liquid fuel, beginning with the activating step of making methanol. This is a particularly pertinent objective based on recent increases in known natural gas reserves; estimates now suggest an abundance of methane gas sufficient to meet current usage for over a century [97]. Technologies that achieve gas-to-liquid fuel (GTL) conversions could leverage these resources, creating end products that coalesce with the existing fuel economy.

Current chemically-based GTL conversion depends upon Fischer-Tropsch synthesis, a method for converting synthesis gas into value-added liquid fuels [164]. Plants based around such conversions face numerous challenges, including high carbon emissions, large capital expenses, and a high level of technological complexity. On top of these difficulties, such processes achieve poor efficiency, typically using only 25-45% of fed carbon substrates and capturing only 30-50% of available energy. These deficiencies pose a substantial obstacle to making chemically-based GTL plants a viable option going forward.

Biological systems present alternative ways to achieve GTL conversions that circumvent many of the obstacles facing chemically-based plants. Methane-consuming organisms, or “methanotrophs”, have been shown to oxidize methane with much greater carbon and energy efficiencies than those realized in Fischer-Tropsch GTL processes [99]. Methanotrophs fall into two categories based on O_2 tolerance:

aerobic bacterial methanotrophs and ANaerobic MEthanotrophic archaea (ANME). Bacterial methanotrophs grow fairly quickly and achieve efficiencies greater than those realized via chemical synthesis, with up to 66.7% carbon efficiency and just over 50% energetic efficiency. However, the ANME are by far the most efficient organisms, theoretically able to convert 100% of available carbon to end products and achieve nearly 80% energetic efficiency. Their remarkable efficiency is tempered by painstakingly slow growth rates, with doubling times as lengthy as 7 months [165].

Due to their exceedingly long doubling times compared with common industrial organisms, ANME organisms remain relatively uncharacterized systems, leaving something of a knowledge gap in this portion of the global carbon cycle [166]. They have thus far eluded efforts to grow them axenically, though several studies have been able to grow them as a highly enriched consortia [165,167]. In spite of culturing difficulties, these limited studies and some culture-independent studies have revealed vital pieces of ANME metabolism. Perhaps most notably, there is evidence that ANME achieve anaerobic methane oxidation (AOM) by performing reverse methanogenesis, using the same set of enzymes found in methanogenic archaea [168]. Considering the clear energetic favorability associated with forward methanogenesis, it is puzzling that ANME are able to subsist and even thrive on a seemingly unfavorable central catabolic process.

Much of the answer to this puzzle appears to come from other organisms present in the ANME consortia, commonly sulfate-reducing bacteria [166]. Several different studies have demonstrated that bacterial partners reduce sulfate or another electron acceptor (e.g. nitrate), essentially coupling the reduction pathway to AOM [166,167,169,170]. The mechanism of electron movement between organisms is not completely understood, though recent publications have shown at least some degree of direct interspecies electron transfer between ANME-bacterial pairs [171,172]. Coupling AOM to a reduction pathway can render the overall transformation energetically favorable under standard

conditions, possibly even more so in specialized environments where ANME consortia thrive, such as deep sea sediments [173]. Thus, the mystery of achieving AOM seems to reduce to performing reverse methanogenesis and donating electrons from methane oxidation into a reduction pathway.

Because of the aforementioned difficulty of growing viable ANME cultures in the lab, these organisms remain poor candidates for laboratory studies, not to mention industrial-scale GTL processes. However, because methanogenic archaea possess the same enzymes as ANME organisms but operate in reverse direction, they represent a viable alternative system for achieving large-scale AOM. Two methanogens in particular—*Methanococcus maripaludis* and *Methanosarcina acetivorans*—have well-developed genetic toolsets [111,174] and have been metabolically reconstructed at genome scale [92]. *M. maripaludis* in particular is an excellent candidate for industrial processes because, although it lacks the methanol utilization genes native to *Methanosarcina* [175], it grows much more rapidly, with doubling time of about 2 hours [109] resulting in ease of making mutations and studying their phenotypes. Moreover, it possesses a much smaller genome with just over 1700 genes compared to 4524 for *M. acetivorans* [110,176], indicating a smaller—and therefore simpler—metabolic reaction network. Based upon these qualities, we selected *M. maripaludis* as our organism of choice for achieving GTL conversion of methane to methanol.

Our strain design strategy for *M. maripaludis* took the form of two basic phases. First, we aimed to insert a pathway for methanol utilization into our organism, enabling conversion of methanol into methane. Having verified growth and methane production on methanol, our objective would become finding a pathway or series of pathways which, when introduced into our methanol-consuming *M. maripaludis*, would enable reverse methanogenesis. To assist in these steps, we turned to iMR540, our genome scale metabolic reconstruction of *M. maripaludis*, as a guide in the metabolic engineering process (for reconstruction details, see Chapter 3). With its dual capabilities for predicting metabolic flux

distributions and overall free energy generation, iMR540 was an ideal platform for generating hypotheses on how to achieve our proposed GTL conversion. This chapter describes the ways in which we used the iMR540 network to explore native hydrogenotrophic methanogenesis and predict ways to alter metabolism for converting methane to methanol.

Methods

In silico Predictions of Genetic Perturbations

In order to predict the effects of genetic perturbations—knockouts, adding non-native pathways, altering growth media—we simulated organism growth using flux balance analysis (FBA) [81] via the Cobra Toolbox 2.0 [132] in MATLAB [7.14.0.739] (The MathWorks Inc., Natick, MA). An in-depth description of using FBA to simulate maximum biomass production can be found in Chapter 3 Methods.

To simulate the effects of an mmp1574 knockout, we utilized the "deleteModelGenes.m" function from the Cobra Toolbox, which restricts flux through affected reactions (i.e. those that rely on mmp1574 to function) to 0. We simulated addition of glycine to the detault *in silico* medium by allowing our model unlimited glycine uptake, which was previously not allowed during growth.

We added non-native reactions using the "addReaction.m" function and specified free energy of formation for any new exchange reactions using values from the eQuilibrator database 2.0 [142]. Whenever possible, we took new reactions directly from the Kbase reaction database and conformed to Kbase nomenclature for metabolite and reaction identifiers. Our scripts for adding the described reduction pathways (see Results) are publically available on Github (https://github.com/marichards/methanococcus).

Estimating Overall Free Energy

As described in Chapter 3, the iMR540 model is configured to generate overall free energy predictions when given standard free energies of formation for exchange metabolites. These standard free energies, taken from the eQuilibrator database 2.0 [142], assume effective metabolite concentrations of 1 mM, pH of 7, and ionic strength of 0.1 M, values that we considered reasonable for microbial systems. For each growth prediction described, we used the "optimizeThermoModel.m" script described in Chapter 3, which calculates overall model free energy and optionally constrains predicted flux distributions to conform to overall $\Delta G \leq 0$.

In all growth simulations prior to predicting reverse methanogenesis, we did not constrain overall ΔG because we were not concerned with the feasibility of these scenarios. For reverse methanogenesis in the absence of additional reduction pathways, we initially predicted growth and overall free energy without constraining free energy, preferring to gauge the magnitude of free energy produced by our predicted solution. We followed this unconstrained scenario by repeating the same simulation while constraining $\Delta G \leq 0$, but were unable to predict growth. Our sensitivity analysis of ΔG vs. equilibrium quotient (Q) for this model was performed without thermodynamic constraints. Instead, we varied Q from 10^{-200} to 10^{0}, plotting the values on a semi-log plot (see Figure 4.5) to determine the value of Q that would fulfill $\Delta G \leq 0$.

For reverse methanogenesis simulations using reduction pathways, we constrained overall $\Delta G \leq 0$ for all simulations. This was necessary because if unconstrained, the model could simply achieve the same solution as it did without the reduction pathways (i.e. it would ignore the new reactions).

Generating a Gene Knockout

Strain MM902, a derivative of *M. maripaludis* S2, was used as the base strain. It contains an in-frame

deletion of uracil phophoribosyltransferase (upt; mmp0680) similar to MM901 [139] for making markerless mutations [177]. However, it also contains ORF1 [178], a plasmid maintenance gene inserted into mmp0680. To construct a knockout of mmp1574, we first amplified the flanking genes with the following primers:

5’ – AAG CGG CCG CAG GTC GTT TGA AAT TTC ATC G – 3’

5’ – AAG GCG CGC CCA TAA AGA CAC CTA ATA AAC AAT C – 3’

5’ – AAG GCG CGC CAT GAT TTA AAC GCT ATT TGT AAC G – 3’

5’ – AAG CGG CCG CTT GAT AAT AAT TAT ATA TAC CC – 3’

We connected these fragments via AscI digestion and sticky-end ligation, then digested and ligated the resulting construct and our vector with NotI. Our chosen plasmid to transform *M. maripaludis* was the suicide vector PCRUptNeoR, a construct indentical to PCRUptNeo [139] except with a deletion of the ampicillin and kanamycin cassette driven by an *E. coli* promoter. After transforming *M. maripaludis,* we selected for the mutant in McCas medium [177] containing 1 mg/mL neomycin. We then selected for cells containing the in-frame deletion of *upt* using medium containing 250 µg/mL 6-azauracil, allowing us to resolve the merodiploid. We sequenced the resulting mutant to confirm our desired genotype, naming the resulting strain “MM1426”.

To test for glycine auxotrophy, we first grew *M. maripaludis* MM1426 cells in McCas medium, which contains casamino acids and therefore supplies glycine. Cells were then subcultured into tubes containing either minimal medium (McNA) or minimal medium supplemented with 10 mM glycine (McNA+Gly). Growth was evaluated via optical density as described previously (Chapter 3) and monitored daily until cell density plateaued.

Results

Predicting Glycine Synthesis

A particularly useful function of a metabolic reconstruction is its ability to elucidate new functions for hypothetical proteins or suggest alternate functions for known genes. Given the abundance of hypothetical proteins identified in the original complete genome sequence for *M. maripaludis* [110], the iMR540 could be an important tool for fully annotating the genome. The reconstruction contains numerous unverified predictions for gene functions, a natural set of hypotheses for biochemical characterization experiments to test their accuracy. By filling out genome annotations more completely, we could work towards mapping out all industrially-relevant pathways in *M. maripaludis*, greatly simplifying the task of harnessing its metabolism to produce desired chemicals. Amino acid biosynthesis, in particular, is an intriguing family of pathways in *M. maripaludis* that would have multiple applications as potential products. Although some of these synthesis pathways—the aspartate/glutamate family, the aromatic family, methionine, alanine—have already been characterized, the majority of these syntheses remain unknown. Hence, there is ample opportunity to use the iMR540 reconstruction to generate hypothesis for these unknown amino acid syntheses.

Glycine is among the group of amino acids for which no biosynthesis pathway is currently known. Unlike most other amino acids, no clear synthesis pathway is predicted by either the MetaCyc or KEGG pathways databases [85,130]. Because glycine is defined as a component of biomass in iMR540, the reconstruction necessarily predicts a pathway for glycine synthesis, depicted in Figure 4.1. As shown, the mmp1574 gene is predicted to code for an enzyme that catalyzes the final step in this pathway, removing an acetyl group from L-2-amino-acetoacetate to synthesize glycine. Utilizing the reaction likelihood scores from our reconstruction, we found this reaction had a probability of 0.39, suggesting that although there was some genetic evidence for its inclusion, there is a large degree of uncertainty

with its place in the network. Notably, our reconstruction also associates mmp1574 with an 8-Amino-7-oxononanoate synthesis reaction at a likelihood of 0.59, a higher value that still shows a fair amount of uncertainty. Thus, this gene could hypothetically code for an enzyme that catalyzes one, both, or neither of these reactions. By characterizing this gene via knockout experiment, we could potentially determine whether or not glycine synthesis is among its functions.

As a first step in evaluating this predicted glycine synthase gene, we used iMR540 to simulate a knockout of mmp1574. We were unable to predict growth of this mutant on H_2 + CO_2 or on formate; however, when our *in silico* medium was supplemented with glycine, the mutant model was predicted to grow under both conditions. Based upon this computational prediction, we hypothesized that knocking out the mmp1574 gene *in vivo* would create a glycine auxotroph. We successfully created this Δmmp1574 mutant *in vivo* (see Methods) and have observed that it grows in medium supplemented with casamino acids. Encouragingly, our mutant strain grows more slowly under these conditions than do wild type cells, possibly suggesting that the mutant cannot grow as quickly due to limited glycine availability. However, results from auxotrophy growth experiments (see Methods) have been inconclusive thus far, therefore we cannot currently determine whether or not our *in silico* prediction is correct. If the Δmmp1574 mutant can grow in both cases, this would suggest that *M. maripaludis* possesses a different glycine synthesis pathway and that mmp1574 may not be linked to any sort of glycine synthase function. However, if the mutant displays glycine auxotrophy, this would suggest that mmp1574 is essential for glycine synthesis, supporting the prediction of our model.

Predicting Methanogenesis from Methanol

Following our exploration of glycine synthesis in native metabolism, we began using our reconstruction to predict novel strain designs. As described in Chapter 3, *M. maripaludis* is a hydrogentrophic methanogen and does not contain the methanol methyl transfer pathway necessary to grow using

methanol. This is in contrast to other methanogens such as the *Methanosarcina*, which can grow using methanol and other 1-C substrates [154]. Prior to achieving our proposed overall conversion of methane to methanol, it was essential to ensure that we could introduce pathways for the reverse reaction; that is, reducing methanol to methane. Thus, our first metabolic engineering task was determining the necessary reactions for uptake and reduction of methanol.

As illustrated by Welander and Metcalf for *Methanosarcina barkeri* [154], methanogenesis from methanol instead of carbon dioxide requires only one additional enzymatic reaction: a methanol:methyl-CoM methyltransferase (Mta). To test whether adding this reaction would enable *M. maripaludis* to catabolize methanol, we added the Mta reaction to the iMR540 model, plus methanol uptake and diffusion reactions that allowed us to introduce methanol into our *in silico* medium. Notably, methanogens are known to grow on both methanol alone and methanol plus H_2 [154]; we configured our model in both growth scenarios to test our model's ability to predict growth under both conditions. We considered our simulation with H_2 as our default case because it presented a less constrained set of growth substrates than did methanol alone. Moreover, we hypothesized that supplying H_2 would be more stoichiometrically favorable as it provides additional electrons that enable virtually all carbon in methanol to be converted to methane, rather than requiring that some carbon be oxidized to CO_2. This is demonstrated by the overall chemical reactions, both with H_2:

$$CH_3OH + H_2 \rightarrow CH_4 + H_2O$$

and without:

$$4\ CH_3OH \rightarrow 3\ CH_4 + CO_2 + 2\ H_2O$$

We simulated our model under both scenarios with no other changes to our reactions and were unable to predict growth in these cases. The cause of these non-growth predictions was our constraint on Eha/Ehb hydrogenase to ≤10% CH_4 flux, a hypothesized flux constraint based on the observed

anaplerotic role of this enzyme (see Chapter 3). Much as in the case of aceticlastic methanogenesis discussed in the previous chapter, methylotrophic methanogenesis skips forward operation of methyl-H_4MPT:CoM methyltransferase (Mtr), the essential Na^+-translocating step of hydrogenotrophic methanogenesis. As illustrated in Figures 4.2 and 4.3, without function of Mtr the only remaining Na^+ pump is Eha/Ehb, which requires reduced ferredoxin in order to create the ion gradient for ATP synthesis. Because the model is no longer supplied with CO_2, the Fwd reaction must generate a small amount of biosynthetic CO_2, somewhat more than is used by the CODH/ACS reaction to synthesize acetyl-CoA. This results in a net gain of reduced ferredoxin and with no other sink for this electron carrier, the Eha/Ehb hydrogenase reaction must operate at high capacity—slightly higher flux than methane secretion—in order to balance ferredoxin usage. With Eha/Ehb flux restricted, reduced ferredoxin cannot fulfill mass balance constraints, rendering the pathway infeasible and resulting in a prediction of non growth.

Assuming that our hypothesized flux constraint is valid, our initial predictions suggested that even with the Mta reaction, methane production from methanol depends on removing the flux constraint from Eha/Ehb hydrogenase. Indeed, when we removed out 10% flux constraint from the model, we could predict growth on methanol both with and without H_2. Our predicted rates and yields for these simulations are shown in Table 4.1 compared with native hydrogenotrophic predictions. Based upon this experiment, we suspect that achieving methanol to methane conversion *in vivo* will likely require supplementing native Eha/Ehb flux, either by overexpressing these enzymes or by adding a similar hydrogenase shown to be capable of supporting methylotrophic growth (e.g. Ech hydrogenase from *M. barkeri* [154]). Efforts to successfully express the Mta genes from *M. acetivorans* into *M. maripaludis* are ongoing; when this process is completed, we expect to characterize the resulting mutant to determine whether or not it can indeed subsist on methanol and H_2 in the absence of CO_2. This anticipated

experiment will not only reveal whether we created our proposed methanol-consuming construct, but also inform the validity of our hypothesized constraint on Eha/Ehb flux.

An added benefit of this simulation is that it gives us a window into central metabolism for a methanol-consuming methanogen without cytochromes, i.e. *Methanosphaera stadtmanae.* As proposed by Thauer [102], energy conservation in this metabolic pathway requires both electron bifurcation by Hdr and sodium gradient formation via ferredoxin oxidation in Eha/Ehb hydrogenase. Though we have essentially added only one internal reaction to the iMR540 model, our resulting reaction map of methanogenesis (see Figure 4.2) looks remarkably similar to that proposed for *M. stadtmanae*. This demonstrates an additional capability of our model as not only a representation of *M. maripaludis* metabolism but also a proxy for other closely-related organisms. There is great potential to leverage this sort of capability for further investigating other organisms, an avenue that is explored more deeply in Chapter 5.

Strain Design for Reverse Methanogenesis

Having successfully predicted methanogenesis from MeOH with the caveat of a fully-functional membrane bound hydrogenase, we turned out attention to the reverse process. As described, the primary known obstacle to reversing methanogenesis is achieving thermodynamic feasibility. Take, for example, the exact reverse reaction of methanogenesis from methanol:

$$CH_4 + H_2O \rightarrow CH_3OH + H_2$$

This reaction produces ΔG = +98.4 ± 10.2 kJ/mol at standard conditions (see Methods) a highly unfavorable overall free energy attributable primarily to consuming water, which has very low free energy of formation (ΔG_f = -157.6 ± 1.6 kJ/mol).

Stoichiometric issues are also a concern, particularly because backwards flux of native pathways could have an adverse effect on conserved pools of electron carriers. As previous efforts have shown, altering the ratios of reducing equivalents such as NAD and NADP in a metabolic network can profoundly impact the model [179,180]. This is particularly poignant for our model of *M. maripaludis,* which contains not only these traditional electron carriers but also several unique compounds such as coenzyme F_{420}. The iMR540 metabolic model provided an excellent way to both examine the stoichiometric feasibility of AOM and determine ways to satisfy thermodynamic constraints.

Stoichiometric Considerations

From examining our methanol-consuming reaction network, we quickly zeroed in on a stoichiometric issue for operating this pathway in reverse. During reverse methanogenesis, we deduced that Hdr must function in reverse to regenerate heterodisulfide, thus oxidizing rather than reducing ferredoxin. Much like methanogenesis from methanol, the reverse methanogenesis pathway from methane to methanol did not activate forward operation of Mtr, requiring EhA/EhB as the primary sodium pump. As depicted in Figure 4.4, EhA/EhB also requires reduced ferredoxin, effectively placing a dual demand on this electron carrier and leaving catabolism without a way to reduce it. Simultaneously, we discovered that our model was predicted to produce a large amount of electron-rich hydrogen through Hdr; thus we hypothesized that we could supply necessary reduced ferredoxin by adding an Fd:H2 oxidoreductase (FHor) to shift electrons from hydrogen to ferredoxin. Indeed, when we added the FHor reaction to our model we found that we could predict flux of methanol from methane (see Figure 4.4).

This effectively completed reverse methanogenesis as far as stoichiometry, but did not yet address our greater challenge of predicting a way to make AOM thermodynamically feasible. As anticipated, our model predicted our methane to methanol conversion to be energetically infeasible, generating +1.37

kJ/gDCW·h to produce 10 mmol/gDCW·h of methanol (or +97.7 kJ/mol CH_4 oxidized). Free energy can also vary considerably with changes in effective intracellular concentrations, grouped into the equilibrium quotient (Q) and related to free energy by:

$$\Delta G = \Delta G^{\circ} + RT \ln Q$$

Given the difficulties of measuring intracellular concentrations, we simulated the effects of concentrations on free energy by performing a sensitivity analysis on predicted free energy based on varying Q. As shown in Figure 4.5, our analysis estimated that our overall chemical transformation would cross into thermodynamically feasible territory ($\Delta G \leq 0$) at $Q \approx 10^{-142}$. This exercise suggested two notable things: (1) free energy of reverse methylotrophic methanogenesis is predicted to be sufficiently endergonic such that even a reasonably steep concentration gradient would not make it a feasible process; (2) given that AOM cannot be driven by manipulating concentrations, our most promising avenue is to temper this unfavorable oxidation process by introducing a favorable reduction pathway to serve as an electron sink.

Energetic Considerations

As described by Mueller et al [99], a variety of electron sinks fill this need in nature, including sulfate, nitrate, and metal oxides. Recently, it was shown that reducing ferric ions to ferrous ions produced sufficient energy to drive AOM from methane to acetate in *M. acetivorans* [181]. Furthermore, the same group predicted that several other reduction pathways could theoretically achieve the same conversion, albeit with less favorable overall free energies [182]. We expected these reduction pathways could also make our proposed methane to methanol conversion feasible, thus we used these same reductions to predict reverse methanogenesis in the iMR540 model (excepting manganese oxide reduction, for which we could not reasonable estimate ΔG_f). The full reaction pathways added for each reduction are shown

in Figure 4.6. Simulations were performed assuming production of 10 mmol/g(dry weight)·h methanol and using thermodynamically constrained FBA, as described in Methods.

In troubleshooting these additional pathways, we immediately encountered a key difference between our organism and *Methanosarcina acetivorans,* the organism used in the previous study. Unlike *M. acetivorans,* which contains cytochromes and therefore has shown propensity for carrying out reducing reactions via membrane-bound complexes, we do not have evidence that *M. maripaludis* is capable of the same transformations. Thus, we assumed for our simulations that any hypothesized reduction pathway would have to proceed within the intercellular compartment. This raised an unanticipated issue with metabolite transport, particularly concerning nitrate reduction. The *M. maripaludis* genome is predicted to code for an ABC type (ATP dependent) nitrate transporter [14], implicating that every mole of nitrate reduced would require hydrolysis of a mole of ATP. Because ATP production is the key factor in synthesizing biomass *in silico,* this predicted transporter had a profound effect on predicted flux distributions that included nitrate reduction. In order to prevent nitrate reduction from effectively using up all intracellular ATP, we had to add either a non-ABC transporter for nitrate or an ABC transporter for nitrite. So long as both nitrate and nitrite possessed the same type of transporter, ATP balance was maintained in predicted flux distributions. With regard to this transporter issue, we ensured that our hypothetical sulfate transporter was not ATP-dependent to match our sulfide transporter; iron transporters were unaffected as both species possess ABC transporters. Though this transporter problem was primarily a mass balance obstacle for our *in silico* growth predictions, it suggested the possibility that transporter types could affect our ability to successfully reverse methanogenesis.

Moving past our choice of transporter types, our simulations revealed that only nitrate reduction was predicted as a feasible way to achieve our proposed conversion. Using either NAD or NADP as an electron carrier, our model predicted that reducing 6.6 mmol/g(dry weight)·h nitrate was sufficient to

offset production of 10 mmol/g(dry weight)·h methanol, assuming standard conditions. To understand the mechanism driving this successful simulation, we compared this predicted flux distribution to that without energetic constraints. Compared to our energetically infeasible solution, our feasible nitrate-utilizing solution predicted only ⅓ the biomass yield, despite approximately the same rate of AOM. This cutback in biomass production was vital for improving energetic feasibility because biosynthesis pathways necessitate influx of CO_2, an energetically expensive reactant. By cutting down CO_2 usage by about 33%, our model greatly improved its energetic outlook. Despite this reduction in growth rate, ATP production did not fall nearly as much, largely in part to the demand from non-growth associated maintenance (NGAM). Because we assume NGAM to be constant for our model regardless of other fluxes, our model must continue to produce extra ATP, subsequently producing extra water as well. Working with a small subset of biosynthetic reactions, our proposed nitrate reduction must necessarily produce water in order to maintain overall water balance withint he model. Thus, our nitrate reduction model was successful because it allowed us to simulate lower growth yield required for energetic feasibility while still maintaining mass balance.

Considering the aforementioned requirements filled by nitrate reduction, it is rather trivial to see why reducing iron was predicted to be unsuccessful. Unlike nitrate reduction, iron reduction produces no water; hence, even though iron reduction is energetically more favorable than nitrate reduction, lack of water in the reduction pathway violated our constraint of mass balance and rendered the network stoichiometrically infeasible. This same issue did not plague sulfate reduction, which produces 3 moles of water per mole of sulfate reduced. Unfortunately, sulfate reduction suffers from too much ATP usage, similar to the problem encountered when using ABC transporters. Because sulfate reduction uses 1 mole of ATP per sulfate reduced as shown in Figure 4.6, it places an enormous demand upon ATP that

competes with that required for biomass formation. Sulfate reduction is predicted to exhaust all ATP available for biomass, resulting in our prediction of no feasible solutions.

Based upon these simulations, our model suggests that nitrate is the only stoichiometrically and energetically feasible pathway that we tested. Coupled together with our previously described FHor reaction, we predict that adding an NAD(P)-dependent nitrate reduction should be sufficient to drive AOM from methane to methanol at standard conditions. Moreover, we could also achieve more favorable conversion by affecting effective metabolite concentrations for major metabolites (methane, methanol, nitrate, nitrite). Though this designs has not yet been achieved in lab, our predictions lay the groundwork for achieving reverse methanogenesis to methanol in *M. maripaludis* in the future and could have important implications for the strategies taken to achieve this conversion.

Discussion

Our uses of the *M. maripaludis* model demonstrate its promise for advancing understanding of methanogenesis and the factors needed to couple growth to reversal of the same pathway. By leveraging the iMR540 model, we made novel inferences about native *M. maripaludis* metabolism and went beyond the parameters of our wild type model to simulate hypothetical metabolic scenarios. The results of these studies allow for an understanding of the metabolic constraints and bottlenecks involved in engineering *M. maripaludis* for gas to liquid conversion.

Our examination of hydrogenotrophic methanogenesis revealed a novel prediction of the essential gene for glycine synthesis, mmp1574. Using this prediction, we hypothesized that knocking out this gene, mmp1574, would result in a glycine auxotroph. Though testing hypothesis remains a work in progress, any outcome of the gene knockout experiment should advance our understanding of *M. maripaludis.* A positive outcome, in which we successfully determine that a Δmmp1574 mutant is a glycine auxotroph,

would support our model's prediction and add this reaction pathway to the library of biochemically characterized reactions in *M. maripaludis*. A negative outcome, in which we determine that the Δmmp1574 mutant can survive without glycine supplementation, would point to an incorrect annotation model, whereby some other mechanism of glycine synthesis exists in the organism. This knowledge would allow us to revise the iMR540 model by finding a different pathway for glycine synthesis. Our glycine auxotrophy experiment is but one example of a myriad of biochemical characterization experiments suggested by the iMR540 model. Numerous gene-associated reactions in the model lack any sort of reference material, originating instead from Kbase annotations. By methodically characterizing many of these reactions, we hope to improve this first model iteration and pave the way for an updated model with better understood biosynthetic pathways.

Going beyond the wild type model, we used iMR540 to predict the pathway of methylotrophic methanogenesis from methanol, mimicking the central catabolic pathway observed in *M. stadtmanae.* Our unsuccessful initial efforts forced us to consider the consequences of our model's constraint on flux through Eha/Ehb hydrogenase and suggested that achieving this conversion may require overexpressing the enzyme or cloning in a more active hydrogenase. When relaxing this flux restriction, we successfully predicted growth from methanol, a result that we are currently working to replicate in lab. The disparity between progress with *in silico* predictions as compared to experimental wet lab work illustrates a drawback of the modeling approach. In our predictions, we innately assume full expression and function of all enzymes in our system, an ideal scenario that rarely plays out in such an efficient manner. Thus, our model predicts the scope of possible fluxes when all enzymes in the network are functioning at sufficiently high expression levels. Actual cloning procedures face many more challenges than are depicted in our model and its predictions, namely difficulties achieving high expression of recombinant proteins in a novel host organism. Although we have not yet achieved the same methanol to methane

conversion *in vivo* that we have shown here *in silico,* our predictions help bound the eventual construct so that we have a better system for troubleshooting our organism once protein expression reaches normal levels.

Predicting methanogenesis from methanol laid the groundwork to reverse the process, finding ways to achieve AOM from methane to methanol. Our initial stoichiometric analysis revealed a deficiency of reduced ferredoxin, a gap that we filled by suggesting a ferredoxin:H_2 oxidoreductase to reduced ferredoxin using electrons from hydrogen. This solved the stoichiometry of reverse methanogenesis to methanol, but was predicted to be energetically unfavorable and therefore infeasible as a final strain design. Using a group of candidate reduction pathways, we predicted that nitrate reduction to nitrite via NAD(P) could successfully offset the positive free energy from reverse methanogenesis, resulting in a design that is predicted to be both thermodynamically and stoichiometrically sound. Other designs resulted in unsuccessful predictions; sulfate reduction was predicted to require too much ATP investment to produce any cell mass, whereas iron reduction was predicted to be stoichiometrically infeasible because of its inability to produce water.

In examining our final reduction path predictions, it is striking to note the differences between our recommendations and those made for *M. acetivorans* by Nazem-Bokaee *et al* [182]. Although they also predicted nitrate reduction as a viable electron sink for reverse methanogenesis, their most viable candidate was iron reduction, a pathway that we predicted to be infeasible. Some of the differences in our predictions can be attributed to different end goals; we produced methanol whereas they produced acetate, an end compound that results in generating 1 mole ATP per mole acetate. Considering that ATP usage is the chief factor preventing sulfate reduction from being effective in our model, it is reasonable to expect that producing acetate instead of methanol would enable us to feasibly use sulfate reduction. However, the majority of discrepancies in our predictions—particularly those linked to using iron

reduction—likely resulted more from differences in overall network topology. The *M. acetivorans* model contains many pathways that are not known to occur in *M. maripaludis,* particularly those involving membrane-associated electron transfers. By rigorously curating our model with non-generic information, we successfully diversified its hydrogenotrophic core metabolism from that of methanogens with cytochromes, like *M. acetivorans.* As a result, we cannot predict the same solutions as were realistic for the *M. acetivorans* model; however, because of the specificity of our model to *M. maripaludis* rather than to a more generic methanogenesis scheme, we have higher confidence in our predictions than if they matched those for *M. acetivorans* exactly.

Overall, our reverse methanogenesis predictions yielded only one feasible strategy of the three different strategies we tested, giving us just one strain design. This scenario is not ideal; instead, we would prefer to have many different possible designs ranked against one another based upon thermodynamic and stoichiometric feasibilities. This result is largely due to our restricted space of possible reduction pathways, which was limited to three different options corresponding to pathways previously observed in methanotrophic archaea. On one hand, limiting ourselves to these strategies lends a measure of confidence to our predicted strain design because this same nitrate reduction pathway has been successful for AOM in other organisms. On the other hand, it is likely that other reduction pathways in the space of known metabolism could also fill the same stoichiometric and energetic needs of our process. As we look to expand our scope of possible strain designs, our path will inevitably lead toward automated methods that leverage reactions in annotation databases. These automated methods could increase our pool of metabolic engineering strategies by moving outside pathways known to partner with methanotrophy in AOM consortia. However, it is vital that as we embrace these methods, we subject each proposed design to the same rigorous standards of feasibility to realistically determine the viability of each strategy.

Tables and Figures

Substrate(s)	$CO_2 + H_2 +$ Acetate	$CH_3OH + H_2$	CH_3OH
CH_4 Secretion Rate	50 mmol/gDCW·h	50 mmol/gDCW·h	50 mmol/gDCW·h
Overall Equation	$CO_2 + 4\ H_2 \rightarrow 2\ H_2O + CH_4$	$CH_3OH + H_2 \rightarrow H_2O + CH_4$	$4\ CH_3OH \rightarrow 2\ H_2O + 3\ CH_4 + CO_2$
Biomass Flux	0.0973	0.0969	0.0969
ΔG (kJ/gDCW)	-5.59	-4.91	-4.43
ATP Yield (per CH_4)	0.55	0.475	0.475
Yield (gDCW/mol CH_4)	2.81	2.79	2.79

Table 4.1: A comparison of predicted parameters for growth of *M. maripaludis* on different media.

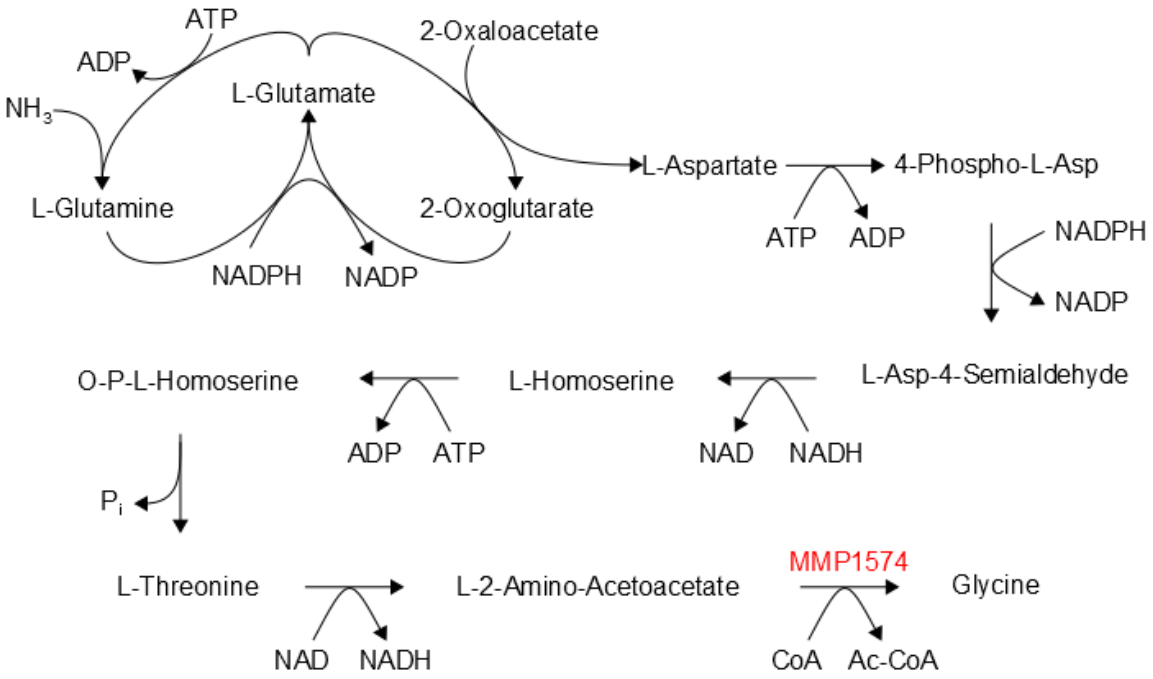

Figure 4.1: Path for glycine synthesis predicted by the iMR540 reconstruction. The mmp1574 gene, shown in **red**, was chosen for knockout based upon this hypothetical pathway. As shown by this reaction scheme, knockout of mmp1574 should prohibit glycine synthesis and render the organism unable to achieve growth. Metabolites: P_i, phosphate; O-P-L-Homoserine, ortho-phosphate-L-homoserine; CoA, coenzyme A; Ac-CoA, acetyl-CoA; 4-Phospho-L-Asp, 4-phospho-L-aspartate; L-Asp-4-Semialdehyde, L-aspartate-4-semialdehyde

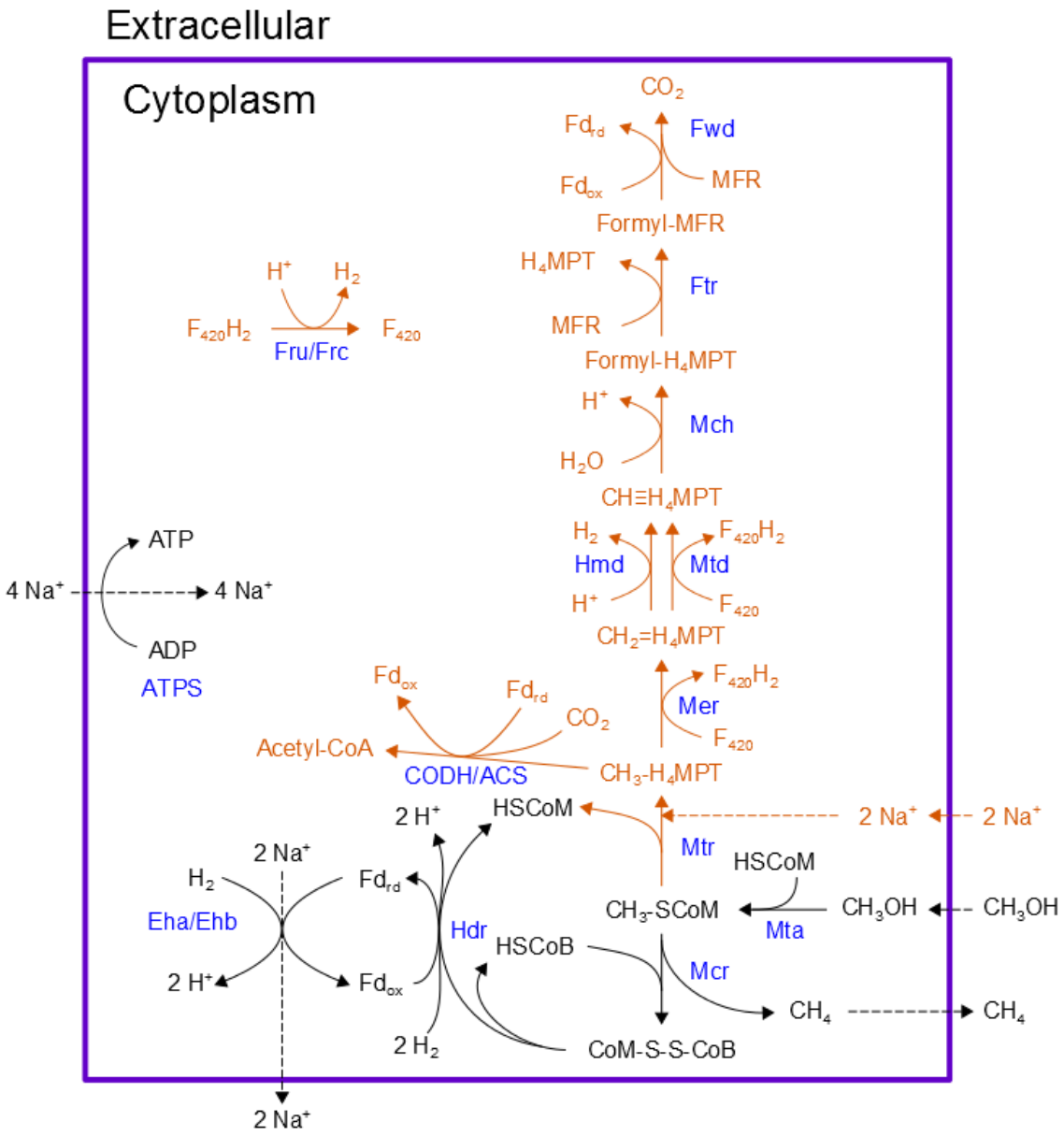

Figure 4.2: Pathway for methanogenesis from methanol and H_2 in *M. maripaludis* as predicted by the iMR540 model. Pathways shown in **black** are main pathways predicted to carry stoichiometric levels of flux; pathways shown in **orange** are secondary pathways predicted to carry small amounts of flux necessary for biosynthesis reactions. Importantly, pathways of the same color do not necessarily carry exactly the same flux. Enzyme names are shown in **blue**. Metabolites: Fd_{rd}, reduced ferredoxin; Fd_{ox}, oxidized ferredoxin ; MFR, methanofuran; HSCoM, coenzyme M; HSCoB, coenzyme B; F_{420}, coenzyme F420. Enzymes: Fwd, formylmethanofuran dehydrogenase; Ftr, formylmethanofuran/H_4MPT formyl transferase; Mch, methenyl-H_4MPT cyclohydrolase; Hmd, H_2-dependent methylene-H_4MPT dehydrogenase; Mtd, F_{420}-dependent methylene-H_4MPT dehydrogenase; Mer, methylene-H_4MPT reductase; Mtr, methyl-H_4MPT coenzyme M methyltransferase; Mcr, methyl coenzyme M reductase; Hdr, heterodisulfide reductase;Eha/Ehb, energy-conserving hydrogenases; ATPS, ATP-synthase; Fru, F_{420}-reducing hydrogenase (selenocysteine-containing); Frc, F_{420}-reducing hydrogenase (cysteine-containing); Mta, methanol:methyl-CoM methyltransferase; CODH/ACS, carbon monoxide dehydrogenase/acetyl-CoA synthase enzyme complex

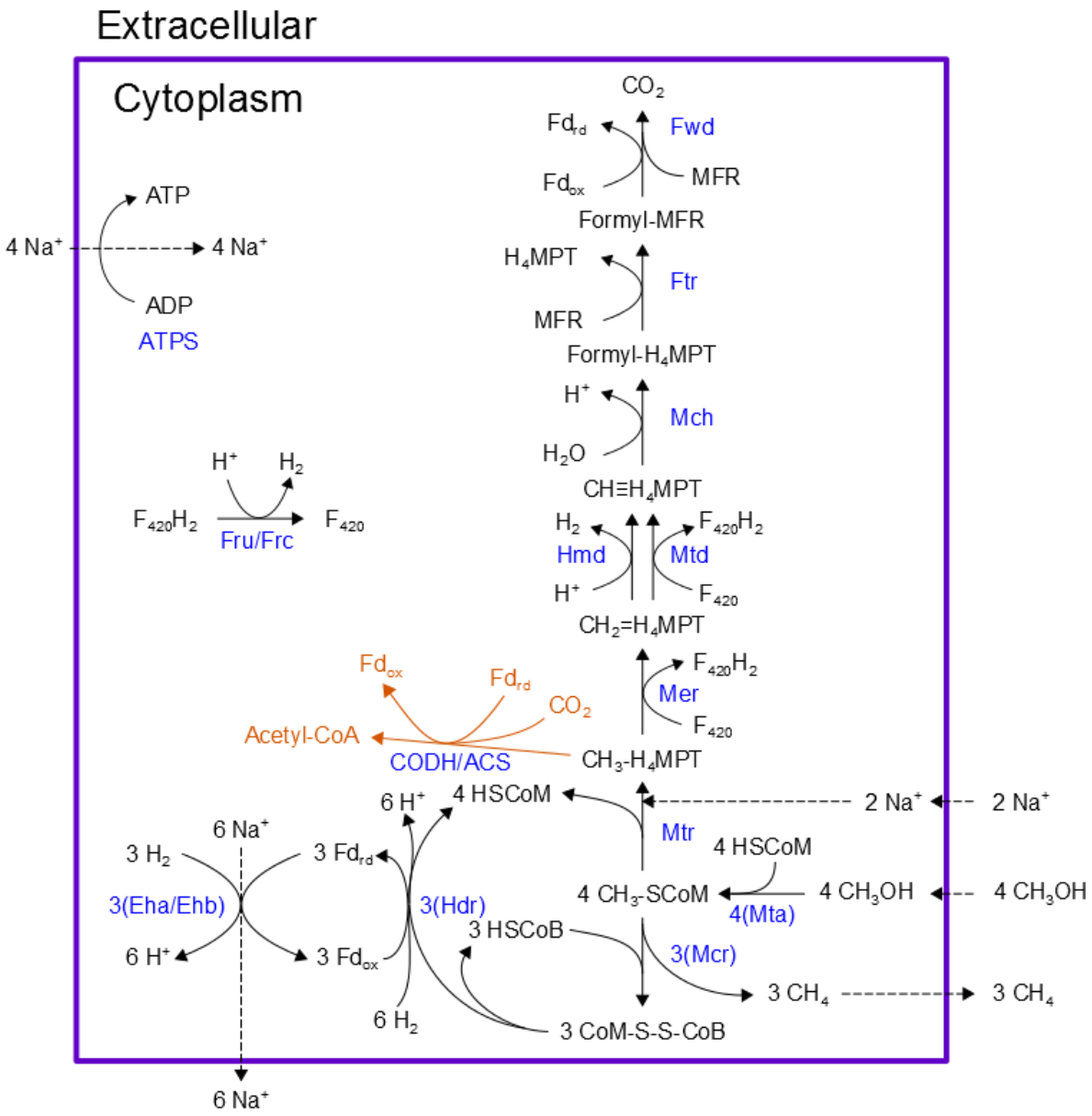

Figure 4.3: Pathway for methanogenesis from methanol alone in *M. maripaludis* as predicted by the iMR540 model. Pathways shown in **black** are main pathways predicted to carry stoichiometric levels of flux; pathways shown in **orange** are secondary pathways predicted to carry small amounts of flux necessary for biosynthesis reactions. Importantly, pathways of the same color do not necessarily carry exactly the same flux. Enzyme names are shown in **blue.** Metabolites: Fd_{rd}, reduced ferredoxin; Fd_{ox}, oxidized ferredoxin ; MFR, methanofuran; HSCoM, coenzyme M; HSCoB, coenzyme B; F_{420}, coenzyme F420. Enzymes: Fwd, formylmethanofuran dehydrogenase; Ftr, formylmethanofuran/H_4MPT formyl transferase; Mch, methenyl-H_4MPT cyclohydrolase; Hmd, H_2-dependent methylene-H_4MPT dehydrogenase; Mtd, F_{420}-dependent methylene-H_4MPT dehydrogenase; Mer, methylene-H_4MPT reductase; Mtr, methyl-H_4MPT coenzyme M methyltransferase; Mcr, methyl coenzyme M reductase; Hdr, heterodisulfide reductase;Eha/Ehb, energy-conserving hydrogenases; ATPS, ATP-synthase; Fru, F_{420}-reducing hydrogenase (selenocysteine-containing); Frc, F_{420}-reducing hydrogenase (cysteine-containing); Mta, methanol:methyl-CoM methyltransferase; CODH/ACS, carbon monoxide dehydrogenase/acetyl-CoA synthase enzyme complex

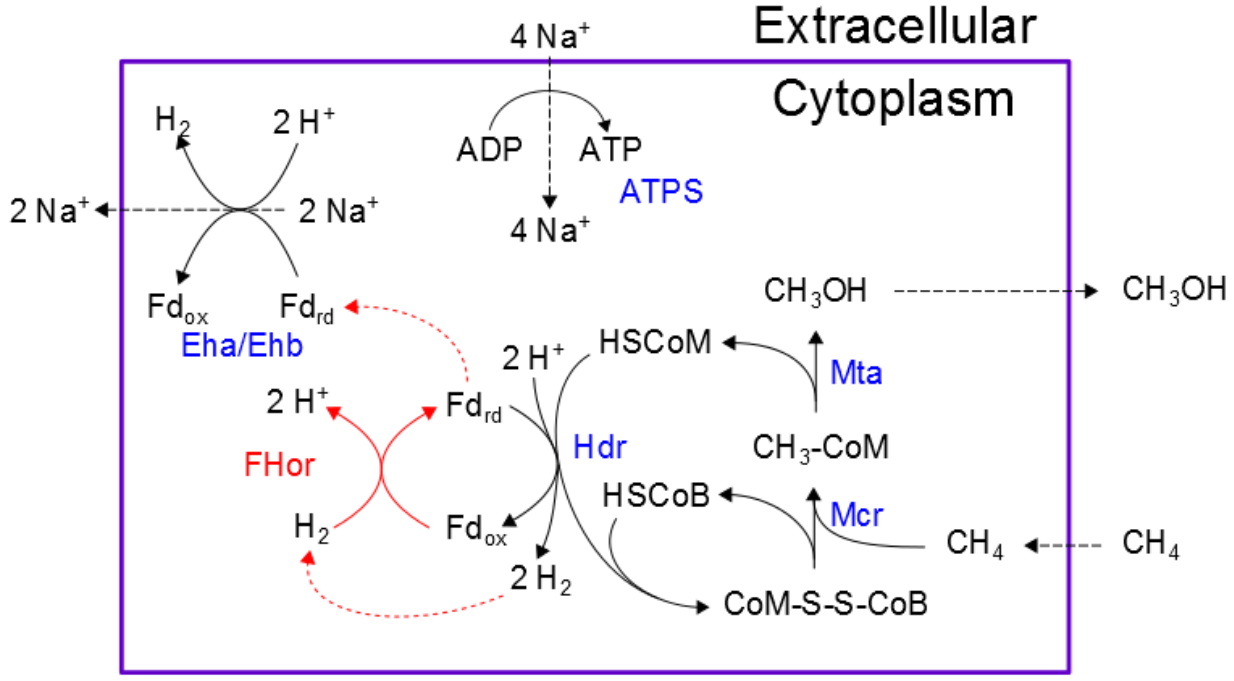

Figure 4.4: Stoichiometric solution for converting methane to methanol in *M. maripaludis* as predicted by the iMR540 model. The added non-native reaction that replenishes reduced ferredoxin using electrons from hydrogen to complete the metabolic pathway is shown in **red**. Enzyme names are shown in **blue.** Metabolites: Fd_{rd}, reduced ferredoxin; Fd_{ox}, oxidized ferredoxin; HSCoM, coenzyme M; HSCoB, coenzyme B. Enzymes: Mcr, methyl coenzyme M reductase; Hdr, heterodisulfide reductase;Eha/Ehb, energy-conserving hydrogenases; ATPS, ATP-synthase; Mta, methanol:methyl-CoM methyltransferase; FHor, ferredoxin:hydrogen oxidoreductase

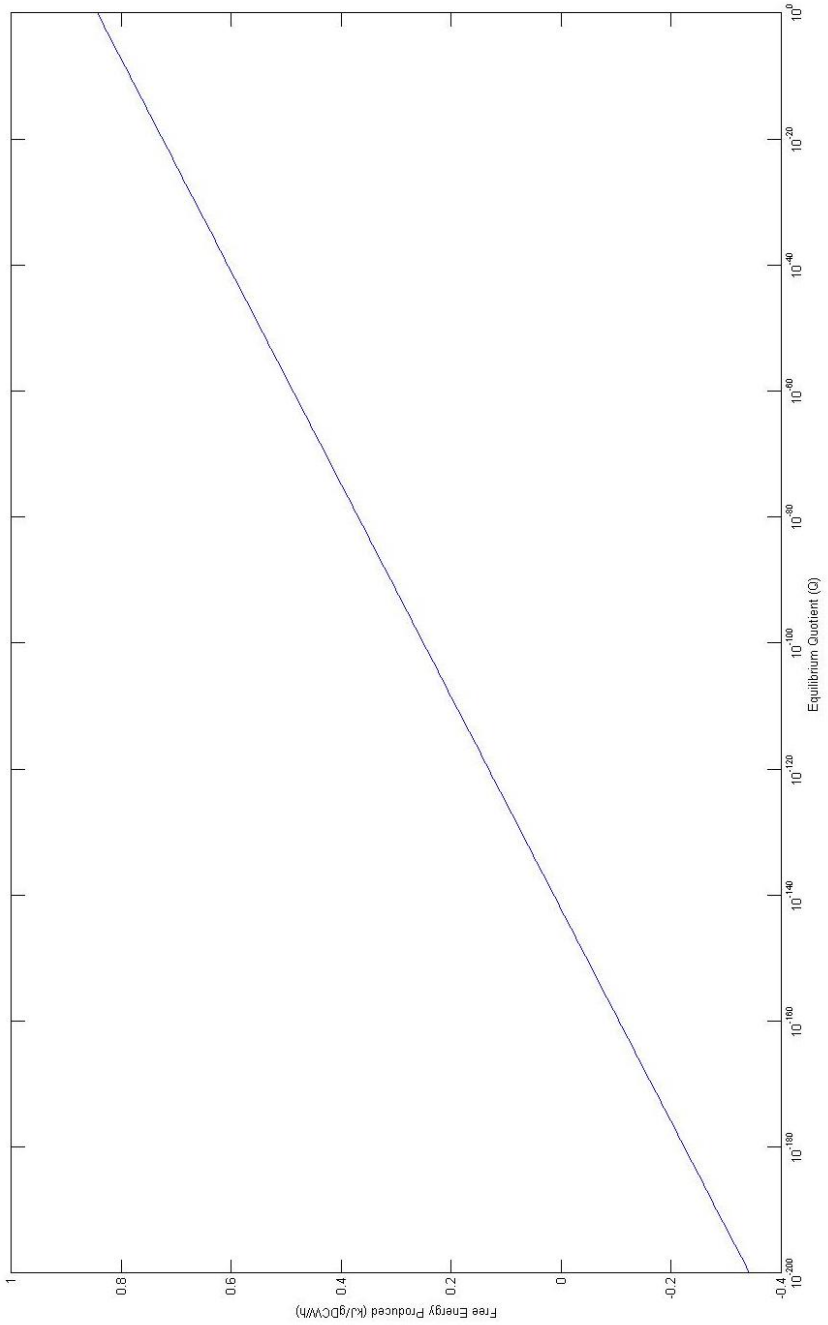

Figure 4.5: Free energy predicted by the iMR540 model as a function of equilibrium quotient, Q, during reverse methanogenesis. These values were calculated in the absence of any additional reduction pathways, aimed at showing the effects of relative metabolite concentrations on free energy generation in the absence of other factors. As shown here, overall free energy generation does not cross into the realm of energetic feasibility ($\Delta G \leq 0$) until $Q \approx 10^{-142}$

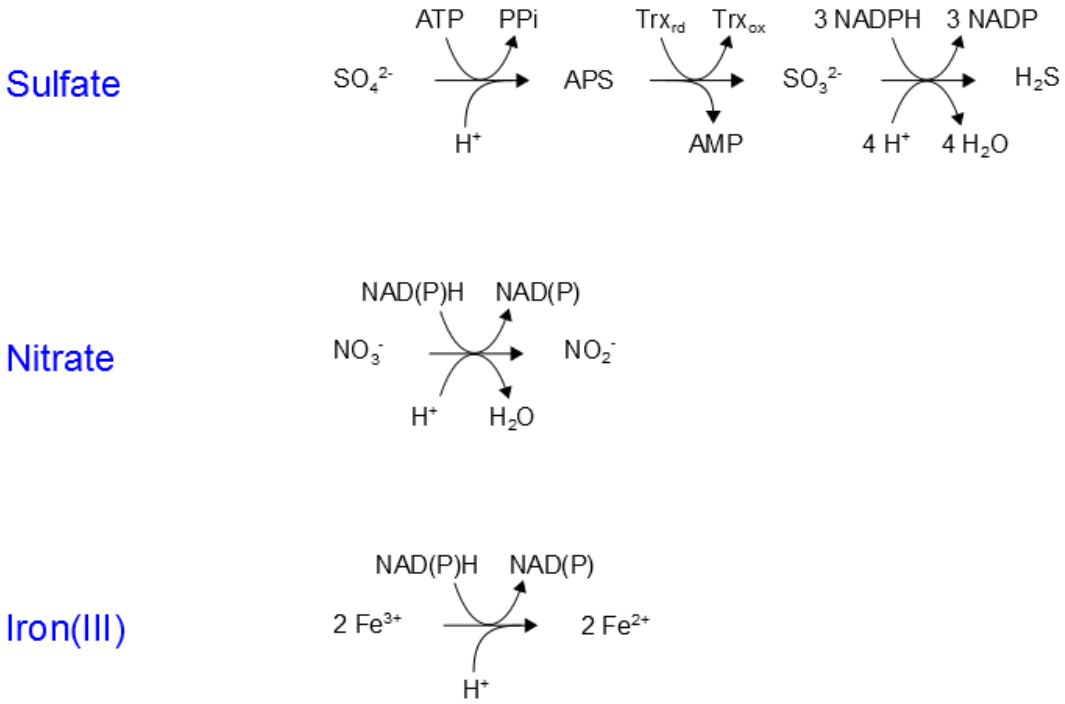

Figure 4.6: Proposed reduction pathways tested in iMR540 to predict reverse methanogenesis. As shown, sulfate and nitrate reduction both produce water but iron reduction does not. Sulfate reduction is also the only pathway of the three that requires 1 mole of ATP per mole of substrate.

Chapter 5: A Method for Perpetuating a Metabolic Reconstruction

Introduction

As extensively covered in Chapter 1, creating a high quality GENRE or GEM is a laborious process, but the resulting model can be well worth the effort. Chapter 4 demonstrated just a few capabilities of a completed GENRE in advancing understanding of native metabolism and guiding metabolic engineering strategies. Applications of such networks far outnumber those discussed in the previous chapter, ranging from contextualizing omics data to discovering novel drug targets to predicting strategies for producing industrial biochemicals [121,183]. These applications can also be extended beyond singular organisms; models encompassing multiple organisms, such as the core metabolic model of *Desulfovibrio vulgaris* and *Methanococcus maripaludis,* open up the usefulness of GEMs to microbial communities [122]. The recent call for more "community systems biology" points to a strong possibility that GENREs and GEMs containing two or more organisms will soon become commonplace, enabling new avenues of discovery [184].

In acknowledging the many applications of metabolic reconstructions, it is also important to note that they are also somewhat limited in so far as a GENRE applies only to the system being modeled. Given the time and effort put into creating a high quality network, it is reasonable to consider how we might better extend these manual efforts to other organisms and systems. As noted in Chapters 1 and 3, a GENRE is never considered a finished product, but rather a living database that must be iteratively updated to reflect new discoveries. One common way to extend a manually curated GENRE is by updating and expanding the existing reconstruction to create a new one. For example, the Palsson Lab has repeatedly updated its reconstruction of *Escherichia coli* by iteratively expanding its genome

annotations [91,161,162,185]. In some cases, multiple existing GENREs for the same organism have been combined to form a consensus model, as was done for *Saccharomyces cerevisiae* [186] and more recently for *Methanosarcina acetivorans* [182]. Updated reconstructions often incrementally improve upon the original network, as shown in a recent comparison of published *S. cerevisiae* models [187]. Thus, using an existing GENRE or GEM as the foundation for an improved model is an excellent way to extend high quality manual curation. However, although the updated model certainly leverages manual curations in a way that extends their impact, it is still confined to the same organism system.

Perhaps the best example of extending a GENRE to another organism is found in model reconciliation, a technique that simultaneously utilizes two manually constructed GEMs to improve one another [188]. The Papin lab, which pioneered this method, used it to align two closely related organisms—*Pseudomonas putida* and *Pseudomonas aeruginosa*—by eliminating discrepancies caused by subjective modeling decisions. By removing these modeling artifacts, such as arbitrarily using NAD in one network and NADP in the other, they illuminated the remaining differences in the models, believed to be actual biological differences between the organisms. This approach demonstrated the possibility of using two high quality manual GEMs to refine one another, provided they are closely related. Its main drawback is that, like creating a new reconstruction from scratch, reconciliation is a manual process that takes months of curation. Furthermore, although it produces two refined models, these models correspond to the same organisms as the starting models. Though model reconciliation certainly has appeal for comparing two models, a technique that uses a high quality metabolic model to create a model for a new organism, rather than refine an existing model, could potentially be even more informative.

The potency of such a technique was demonstrated by ITEP, a suite of tools for examining microbial pangenomes [189]. The ITEP toolkit includes a tool that takes a GENRE and extends it to related organisms via proteome comparisons, creating partial networks for a clade of organisms based off one

source network. This method powerfully leverages manually curated information in the source GENRE, creating a high quality draft foundation that can be used to compare gene functions between organisms. ITEP itself is not designed as a metabolic modeling platform; rather, it is geared toward studying patterns of gene associations across pangenomes. But a similar approach could be quite effective for extending the benefits of manual curation to a new organism without expending the usual time and effort.

A suitable modeling platform for developing such an approach is the Department of Energy Systems Biology Knowledgebase (Kbase; www.kbase.us). Built on top of the Model SEED database, the Kbase is a collection of data and methods including the Model SEED automated reconstruction algorithm [51]. As described in Chapter 1, automated reconstruction annotates a supplied genome using the RAST (Rapid Annotations using Subsystems Technology) Server [190] to create a draft reconstruction and gap-fills the network to create a GEM that can simulate growth on specified media. A major advantage of using the Kbase is that it supports gap-filling using probabilistic annotation (ProbAnno), a technique that assigns reaction likelihoods based on gene homology and synteny, then gap-fills by maximizing reaction pathway likelihoods [126]. Thus, ProbAnno can predict additional gene annotations for a reconstruction, maximizing the information gleaned from the genome sequence, as was demonstrated in Chapter 3. Taken together, these methods taken together form perhaps the best available model building tool to drive development of new modeling methods.

A method that combines the deployment of the same type of proteome comparison strategy practiced in ITEP with the tools available in Kbase could harness the strengths of both platforms. Given a high quality GEM created through manual curation and the genome of a closely related organism, such a method could pull manually curated information from the source model and supplement it with automated annotations to create a high quality draft model of the related organism. This approach

could mitigate the effects of database misannotations by relying heavily on the manually curated model to inform annotations in the draft model. Simultaneously, such drafts would not be limited to partial network reconstructions but rather, could be extended to full genome-scale draft models using automated reconstruction and ProbAnno gap-filling. Thus, the output of this model morphing method would be a fully functional draft GEM for a previously un-modeled organism. This chapter describes the development of a methodology that attempts to achieve the aforementioned goal: leveraging the biochemical information present in a manually constructed metabolic model to create high quality draft genome scale models of related organisms.

Methods

Method Overview

Scripts for our morphing method were written entirely using the Python programming language. We also used multiple functions in Kbase, including the built-in flux balance analysis (FBA) solver for simulating model growth. For a full description of FBA, refer to Chapter 3 Methods. The Kbase tools described below are contained within the Kbase Application Programming Interface (Kbase API; https://kbase.us). All genomes used for our method were also taken from the Kbase repository.

Proteome Comparisons

The “Compare Two Proteomes” tool compares two entire proteomes to one another based on BLAST output [37,191]. Using the BLAST hits, it finds bidirectional best hits for genes in both proteomes, allowing for many-to-many mappings between genes. The method requires input of two different genome objects and outputs a proteome comparison object containing the results of the analysis. Though the “Compare Two Proteomes” method contains optional parameters that allow users to tune the threshold for matches, here we ran it with default parameters for our purposes.

Automated Reconstruction

We used the "Build Metabolic Model" tool to construct our automated Kbase reconstruction models. The only required input for this tool is a Kbase genome, from which the model builds a draft network using gene annotations. The method optionally allows users to gap-fill the draft network to create a model that simulates growth. For this added functionality, the user must also specify a media formulation for gap-filling. Models are automatically assigned one of four default biomass compositions: Gram positive microbe, Gram negative microbe, Core pathways microbe, or Plant. Optionally, the user can specify one of these biomass objective functions for the final model. The biomass composition affects the gap-filling process because it determines what metabolites are necessarily synthesized to achieve model growth. For our automated reconstructions, we use the default biomass assigned by Kbase. This is distinct from our translated and morphed models, which use the same biomass as our manual model.

Probabilistic Gapfilling

We used the ProbAnno method for likelihood based gap-filling to add reactions with high likelihood to our morphed models and to gap-fill on new media formulations. ProbAnno is based upon gene homology; it estimates the likelihood of multiple possible gene annotations for each gene in the supplied genome, then maps these genes to reactions using the annotations in Kbase. Thus, not every reaction in the database is assigned a likelihood, but those that do scale from 0-1 (1 being most likely). The gap-filling process uses a penalty function to incorporate likelihoods, with smaller penalties for high likelihoods to increase the chances of including these reactions in the final model. A more in-depth overview of ProbAnno is provided by the method's authors, Benedict *et al* [126]. Notably, the ProbAnno method is not currently supported by the Kbase Narrative Interface and must be run through the Kbase API.

Results

Workflow Description

Our final method, shown in Figure 5.1, requires a manually curated model (Model A), a complete genome for that organism (Genome A), a complete genome for a related organism (Genome B), and a growth medium for the related organism (Media B). The two genomes are fed into the Kbase "Compare Two Proteomes" tool, yielding a map of proteins from Genome A onto Genome B. This essentially predicts which genes in Organism A are also present in Organism B and, conversely, which genes in Organism A do not have homologs in Organism B ("A not B" genes). Using the GPRs in Model A together with our proteome mapping, we flag reactions that are likely to also appear in Model B, creating "Translated Model B" (see Figure 5.2). Notably, we do not discard the remainder of Model A in this step; these other reactions are carried over into the "Super Model B" construct, but are flagged either as gap-filling reaction s(i.e. they lack genes in either organism) or as "A not B" reactions. We consider the latter group to be the candidates most likely to reflect biochemical differences between the organisms.

Regardless of phylogenetic similarity, we expect Organism B to contain features that are not present in Organism A; hence, we also feed Genome B into the Kbase "Build Metabolic Model" tool, which returns an automated draft reconstruction of Organism B ("Reconstruction B"). Reconstruction B is also fed into Super Model B to add genomic information that was not present in Model A, increasing our total pool of candidate reactions. We also feed Genome B into the ProbAnno tool to create a list of reaction probabilities, P ($0 \leq P \leq 1$), for Organism B. This allows us to identify reactions missed by Reconstruction B that are likely to be present in Organism B and add these reactions to Super Model B.

Due to differences between GPRs and Reconstuction B and Translated Model B, many shared reactions contain dissimilar gene rules, a discrepancy that must be resolved when merging them to create Super Model B. We resolved these discrepancies using OR relationships, which minimize the constraints

imposed on the final model. Examples of implementing these rules are shown in Table 5.1. In addition to resolving the GPR discrepancies as shown, we also flagged these reactions as high priority targets for manual curation. We created a list of these conflicting GPR reactions, thereby directing the user toward specific pieces of the model that require special attention. This flagging process makes manual curation more directed and streamlined by better prioritizing reactions and genes most likely to require revisions.

After assembling our Super Model B structure, we enforce specific growth conditions for Organism B to ensure that our final morphed model can predict growth. Although biomass composition affects growth predictions and can vary considerably between organisms, we assume that Organisms A and B are sufficiently similar that their biomass compositions can be assumed the same unless measured biomass composition is available for the target organism. Media composition is also quite organism-dependent and despite the fact that many organisms have not yet been cultured, most well-characterized organisms possess defined media conditions. In cases where media composition is unknown, the Kbase contains many generic media conditions that can be used to simulate growth. External databases, such as MediaDB [46] and the KOMODO database [47], also compile known media for sequenced organisms and can be used to select a suitable medium for predicting growth. Using Media B, we gap-filled Super Model B using ProbAnno gap-filling to create Gap-filled Model B. Notably, if Media B is identical or sufficiently similar to the growth conditions used to successfully simulate growth of Model A (Media A), Super Model B will already be able to predict growth. However, if Media B is sufficiently different from Media A, the Super Model B network will contain gaps that prevent successful simulation of biomass formation.

Up to this point in the process, no reactions have been removed from the model, thus Gap-filled Model B contains many more components than any of its source materials. At this juncture, we begin testing candidate reactions for removal from the model by testing each reaction without a gene (gap-fill

reactions). We test essentiality of these reactions via single reaction deletion, working through all gap-fill reactions in order of priority. The priority list is a necessary component of the method because removal order could impact reaction essentiality; reactions tested for removal later are more likely to be essential because degenerate pathways are more likely to have been removed earlier in the process. We begin the list with reactions in Model A that were identified as unlikely to be present in Model B—reactions classified as "A not B"—because these reactions could potentially represent actual functional differences between the two organisms. These are followed by the remaining list of reactions that lack genes in both models, ranked in order of increasing likelihood. Once we test all gap-fill reaction essentiality and remove unessential gap-fills, we arrive at the final product of our method: Morphed Model B (Figure 5.1).

Application to Methanogenic Archaea

To demonstrate the efficacy of our method, we required a group of related organisms with at least one high quality manual model; the methanogenic archaea or "methanogens" are a good candidate for this purpose. In Chapter 3, I reconstructed a model of *Methanococcus maripaludis* based mostly on Kbase annotations and identifiers, allowing us to sidestep some of the nomenclature and formatting issues described in Chapter 1 that would be magnified if we used a model reconstructed from a different database. The iMR540 model of *M. maripaludis* includes considerable manual curation, providing a sizeable candidate list of manually added reactions that offer depth on top of Kbase annotations. Though methanogens are less studied than many bacterial clades, a number of methanogens have been fully sequenced and characterized to some extent because of the interest in tapping into their unique methane-producing metabolisms.

In addition to *M. maripaludis,* we chose three candidate organisms for testing our method. One obvious candidate was *Methanocaldococcus jannaschii*, a hydrogenotrophic methanogen like *M. maripaludis*

that has been used for numerous biochemical characterization experiments. Indeed, much of the literature based information included in iMR540 was linked to studies in *M. jannaschii*. For this reason, morphing our manual model to a draft of *M. jannaschii* presented an excellent opportunity to observe how much of that manual information we could automatically transfer to the new model. We also selected *Methanosphaera stadtmanae*, a somewhat more distant relative to the first two organisms as per Garcia *et al* [192] that consumes methanol and H_2 rather than CO_2 and H_2. It also possesses a smaller genome than the other two methanogens, giving us an opportunity to study effects from phylogenetic difference, growth conditions, and genome size. Finally, we chose *Methanosarcina barkeri,* a much more distant relative of the other three organisms with a far larger genome and large range of possible growth substrates. Notably, *M. barkeri* can achieve growth on CO_2 and H_2, enabling us to study the effects of genome size and phylogenetic distance on our method while controlling for media.

Finally, we also used our method to morph *M. maripaludis* itself, giving us something akin to a control case. From reconstructing the iMR540 model, we understood that it incorporated many gene annotations and reactions that were absent from the Kbase automated model. We were curious to observe the overlap in the models and compare this to what we saw for the other organisms. We were somewhat concerned that by greedily compiling information to create our Super Model B structure, we might run the risk of retaining too much information, a problem that is difficult to assess when creating a model for a different organism. This control afforded us the opportunity to ensure that our methods were not overly greedy and that we erred more toward capturing biological differences between organisms than grabbing all possible information. To assess these effects, we created an automatically reconstructed model for each morphed organism using the default Kbase “Build a Metabolic Model” tool. Each set of organisms ended up with 3 different models for comparison: (1) Morphed Model B, the

final result of our method; (2) Reconstruction B, the automated Kbase model; (3) Model A the iMR540 *M. maripaludis* model.

Morphing to New Organisms

As a first measure of assessing our morphed models, we compared the features in the finished morphs with the other two models for each organism (Figures 5.3-5.5). Looking first at *M. jannaschii* (Figure 5.3), we observed that the final morph contained more genes than either the automated model or the iMR540 model. This seemed to indicate a combination of the larger genome of *M. jannaschii* compared to our source organism and their similarity; the latter would result in many genes being matched to iMR540, whereas the former would explain the slight increase. Examining the features of *M. stadtmanae* appeared to confirm this trend (Figure 5.4), as the final morph contained fewer genes than did iMR540. The smaller genome and greater phylogenetic distance associated with *M. stadtmanae* could account for this observation, as fewer genes would mean less information to add to the morph and less overlap with *M. maripaludis* would mean fewer genes taken from iMR540. Our final candidate, *M. barkeri,* ended up with many more genes than either of the other two morphs (Figure 5.5) despite its greater phylogenetic distance from *M. maripaludis*. Though this seemingly resulted in matching fewer genes with iMR540, the much larger genome size meant that many more gene annotations were added from the automated Kbase reconstruction. Across all three organism morphs, we observed a sizeable jump in the number of metabolites when compared to the other two models, but no corresponding gain in number of reactions. This trend suggests that the morphing process likely removed much network degeneracy, wherein multiple pathways can achieve the same essential chemical transformation. In these cases, where there are N gap-fill reactions to convert one metabolite to another in a necessary pathway step, our essentiality requirement would remove N-1 reactions, leaving only the final reaction tested. This demonstrates the importance of sorting our priority list of reactions to remove; by ordering

our list such that high likelihood reactions are tested last, we increase our chances of keeping these reactions in our model.

Our overviews of model features were somewhat informative, but did not fully explain the trends we saw in gene numbers. We were unsure whether a larger genome size necessarily meant having more genes from iMR540, or whether the final number of genes was more independent from genome size. To better understand the trend, we examined the gene origins for each organism's morph, shown in Figures 5.6-5.8. As demonstrated by Figure 5.6, the morph for *M. jannaschii* takes about 60% of its genes from iMR540, though only a small portion of those genes intersect those from the Kbase model. The *M. stadtmanae* morph (Figure 5.7) displayed some slight differences, with just under 50% of its genes coming from iMR540 and 65% coming from the Kbase model but with similarly poor overlap between the two source models. Somewhat strikingly, the *M. stadtmanae* Kbase reconstruction contained more genes than did the *M. jannaschii* Kbase reconstruction despite having a smaller genome, an observation that was not consistent with our morphed models. The *M. barkeri* morph (Figure 5.8) took only 42% of its final genes from iMR540, compared to 69% from the Kbase model. Its total model contained the smallest proportion of genes from *M. maripaludis*, though it matched more genes in sheer number than did *M. stadtmanae*. This points to the idea that genome size played a role in the final morph in so far as *M. barkeri* had more proteome matches with *M. maripaludis* simply because its larger genome gave it more matching opportunities. Perhaps most importantly, phylogenetic distance seemed to play a role in the portion of the final model that matched the source model. Greater similarity lent itself to a higher percentage of matching genes in the final model and greater distance between organisms resulted in more influence from the automated Kbase model. Also quite telling was the exceedingly poor overlap between iMR540 matches and Kbase genes, which ranged from 10-13%. This observation was

particularly surprising for the *M. jannaschii* morph, which we expected would show some similarity between genes from automated reconstruction and those added from our manual model.

As a final evaluation of our morphs, we examined the reactions in all 3 model types for each organism (see Figures 5.9-5.11). Much as we saw in the previous analyses, *M. jannaschii* showed a bias toward the iMR540 model, sharing the vast majority of its reactions with the manual model. This included nearly 3 times as many reactions from the manual model that were absent from the automated Kbase model as there were in the reverse case. Notably, the 3 different model types still shared a core of about 350 reactions, the majority of the reactions in the morph. As we expected, *M. stadtmanae* did not show nearly as much favor toward iMR540. Although it shared a similarly sized core of about 350 reactions with both source models and a similar number of reactions with just the manual model, it borrowed much more equally from the manual and automated models. Furthermore, fewer reactions taken from the manual model were associated with genes in the final morph, whereas more reactions from the Kbase model were gene-associated (Table 5.2). This same trend was amplified in the *M. barkeri* model, which again shared the same core of about 350 reactions but favored the Kbase model slightly over the manual model. Just like *M. stadtmanae*, the *M. barkeri* morph took many more gene-associated reactions from Kbase alone than it did from the manual model. These observations seem to point again to the increased influence of automated gene annotations with increased phylogenetic distance.

Studying reaction overlap for all models at once, we saw that all morphed models borrowed fairly equally from the manual model, but the *M. stadtmanae* and *M. barkeri* morphs took many fewer gene-associated reactions and had to supplement much more using Kbase. We also noticed the gap-filling process ended up adding 15 reactions unique to the morph for both *M. jannaschii* and *M. stadtmanae*, which were grown on slightly different media from iMR540. This is distinct from *M. barkeri*, which we simulated on the same media as used in iMR540 and thus required no gap-filling. In any case, the gap-

filling reactions represent a very small portion of the other two models and played a very minor role in the morphing process. Conversely, our models all removed approximately 20 reactions that overlapped between both the Kbase and manual models. Compared to the core of about 350 reactions shared by all 3 models for each organism, these ~20 reactions comprise a small subset, showing that we removed very little of the reaction information that agreed between the two source models.

Morphing to M. maripaludis

We performed the same set of analyses for our "control" scenario, shown in Figures 5.12-5.14. Despite morphing our *M. maripaludis* model to itself, a process we expected would change the manual model very little, the features in this experiment showed very similar patterns to the other three cases. We were surprised to observe that like *M. barkeri,* our *M. maripaludis* morph added many new genes to iMR540, resulting in over 700 genes in the final morph. As demonstrated by Figure 5.13, the reason for this large gene increase was poor gene overlap between iMR540 and the Kbase model; only 83 genes, about 11% of those in the morph, occurred in both models. Although nearly 70% of the final morph matched the manual model—easily the highest percentage for any of the morphs—the remaining 30% represented a sizeable set of genes in Kbase that did not occur in our model. This discrepancy reinforced the notion that the manual curation required to create iMR540 added much biochemical detail not captured by Kbase and discarded many parts of the Kbase reconstruction that lacked sufficient evidence. Based on our experience assembling iMR540, which began with the Kbase reconstruction itself, we think most of these additional genes represent mis-annotations and that in this case, the morphing process has somewhat muddled the final morph by including these genes.

Reactions and metabolites also show similar patterns to the other morphed organisms, with a modest increase in metabolites accompanied by a slight decrease in reactions when comparing the morph to the manual model. In examining reaction overlap (Figure 5.14), we saw a continuation of the relationship

between phylogenetic distance and morph reliance on the Kbase automated model. Indeed, when morphing the *M. maripaludis* model to itself, we counted only 37 reactions from the automated model that did not occur in iMR540, the lowest number of these reactions for any morph. Among the reactions taken from only iMR540, this morph also had the highest percentage of gene-associated reactions by far, unsurprising considering that it matched all genes in the manual model. Like the other morphs, this one also contained a large reaction core shared between all three model types, though slightly larger than the others at 392 reactions. Interestingly, our morph discarded 56 gap-filling reactions from the manual model, presumably replacing them with higher likelihood reactions from the Kbase model. These replacements may be an artifact of removal order, particularly because some reactions in iMR540 lack any likelihood scores or genes and are therefore among the first reactions tested for removal.

Discussion

With our novel model morphing method, we have created a new way to extend a high-quality manual reconstruction to closely related organisms. By basing our final model on a proteome comparison with the manual source model, we can capture much information that would likely be missed by an automated draft reconstruction. Our method could have applications not only morphing one organism to another but also in larger scale clade reconstruction, whereby one high-quality seed model could be transformed into a collection of related models and enable pangenomic analyses.

As we demonstrated by morphing the iMR540 model to three functionally-related organisms, our method is able to reflect similarity between organisms. Our morphed model for *M. jannaschii,* a close relative of *M. maripaludis*, displayed much more consistency with the source model than either of the other morphed models, both of which were more distant relatives. Yet, we caution that regardless of phylogenetic or functional distance between organisms, our final morphs are still only draft models. Although our morphing process should drastically reduce the time needed to locate reaction pathway

information compared to starting with an automatically reconstructed model, our morphs still contain a fair amount of redundancy, particularly for reaction GPRs with disagreements between source model and Kbase annotations. Accordingly, we copiously flagged reactions in the final morphs to direct modelers beginning with these drafts toward the areas in most need of manual curation. Based upon our own experience with metabolic reconstructions, these outputs should be a boon to model creation, cutting down significantly on manual curation time.

The need for continuing to employ manual curation is underscored by morphing *M. maripaludis* itself, an exercise that highlighted the differences between our source model and the Kbase automated version. Overlap between these models was much smaller than reasonably expected, a difference that manifested in a final morph with many more genes than either of these two contributing models. It is quite telling that despite the manual model possessing 528 genes and the automated model possessing 322 genes, they share only 83 of them, or 15.7% and 25.8% of their respective gene totals. Though it is possible that some genes from the Kbase model could represent information missed by the manual model, many of them were deliberately removed during manual curation, as described in Chapter 3. This calls into question the ubiquity of the automated reconstruction process, which takes gene annotations from a centralized database, much of which is based off well known pathways in model organisms. For an organism like the methanogens used in this study, it is quite possible that these generalized annotations create draft models dominated by misinformation and could end up slowing manual curation efforts. If indeed an automatically reconstructed draft model can only correctly capture in the range of 10-20% of genes present in a high quality model, it could be folly to use such a draft model, as the small gain in correctly annotated genes could be greatly muddled by the larger gain in mis-annotations. These potential shortcomings further point to the need for a method such as the one presented here to extend the benefits reaped from manual model curation and use that information as

the basis for a new high quality reconstruction rather than relying on automated reconstruction alone. Thus, our model morphing tool could provide a viable alternative for organisms that are dissimilar to those that populate annotation databases, provided a high quality manual model of a closer relative already exists.

On a related note, we do not recommend applying our morphing method for organisms outside a reasonable phylogenetic distance. Just in our small set of organisms, we observed a large difference when morphing to our most distant organism, *M. barkeri,* compared to the other organisms. The final morph relied more on Kbase annotations than on those from the manual model, and we would expect this trend to continue with increased phylogenetic and functional distance. Correspondingly, as our target organism for morphing becomes more dissimilar to our source organism, we are concerned that annotations gleaned from the manual model may supply misinformation in the resulting morph. Attempting to morph a manual model to an essentially unrelated organism could result in numerous misannotations. These annotation errors may greatly cloud any useful information gleaned from the morphing process, resulting in a draft model that hinders the reconstruction process. Instead, model morphing should only be performed using a manual model from a close relative, otherwise, in cases where no such model exists, an automated reconstruction method would likely prove most efficacious. As with any automated method in metabolic modeling, it is vital to exercise caution when using model morphing and to meticulously examine the resulting product to ensure that its quality is satisfactory.

Tables and Figures

GPR Discrepancy Type	Translated Model B	Reconstruction B	Super Model B
Single Gene Disagreement	G1	G2	G1 OR G2
OR Subset Disagreement	G1 OR G2	G1	G1 OR G2
Multiple OR Disagreement	G1 OR G2	G2 OR G3	G1 OR G2 OR G3
AND Subset Disagreement	G1 AND G2	G1	(G1 AND G2) OR G1
Multiple AND Disagreement	G1 AND G2	G2 AND G3	(G1 AND G2) OR (G2 AND G3)
Complex Mixed Disagreement	(G1 AND G2) OR G3	(G2 AND G3) OR G4	(G1 AND G2) OR G3 OR (G2 AND G3) OR G4

Table 5.1: Examples of GPR discrepancy reconciliation from combining the Translated Model B construct and Reconstruction B construct to form Super Model B. These examples display what we consider to be the full spectrum of different cases. Any GPR we encountered could be expressed as a combination of these examples and resolved in the Super Model B structure based on the rules outlined here

	Total Reactions	**% Reactions with Genes**
***M. jannaschii* Morph**	631	86%
iMR540	190	78%
Kbase	70	79%
Both	356	95%
***M. stadtmanae* Morph**	696	81%
iMR540	185	59%
Kbase	146	90%
Both	354	92%
***M. barkeri* Morph**	741	87%
iMR540	183	67%
Kbase	198	94%
Both	359	94%
***M. maripaludis* Morph**	597	93%
iMR540	167	90%
Kbase	37	68%
Both	392	97%

Table 5.2: The reaction origins for all of our final morphs. Percentages reflect the fraction of reactions from each source model that are gene-associated (e.g. 190 reactions in the *M. jannaschii* morph came only from iMR540 and of those 190, 78% of them were associated with at least one gene)

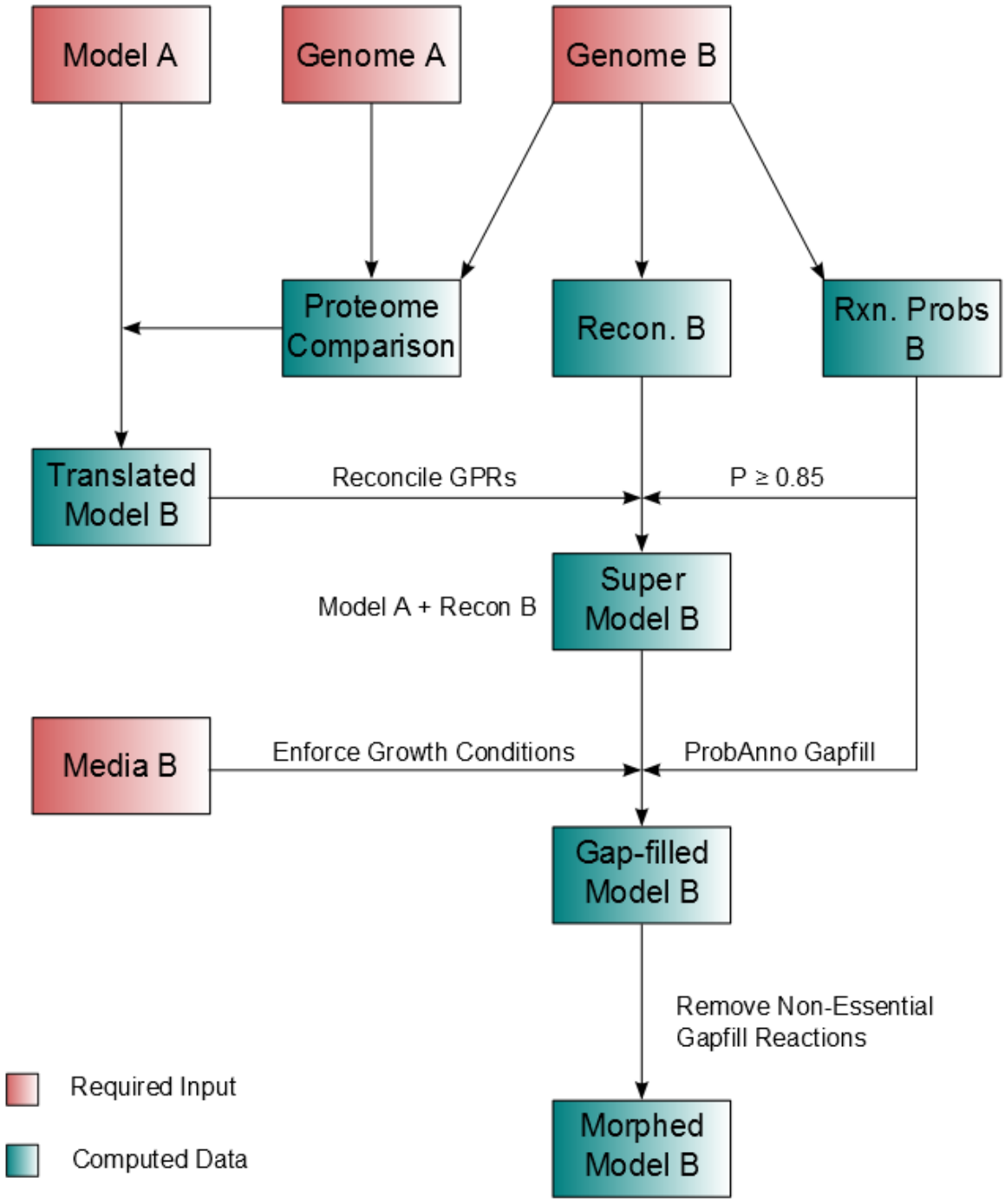

Figure 5.1: A flowchart showing the basic workflow of our model morphing method.

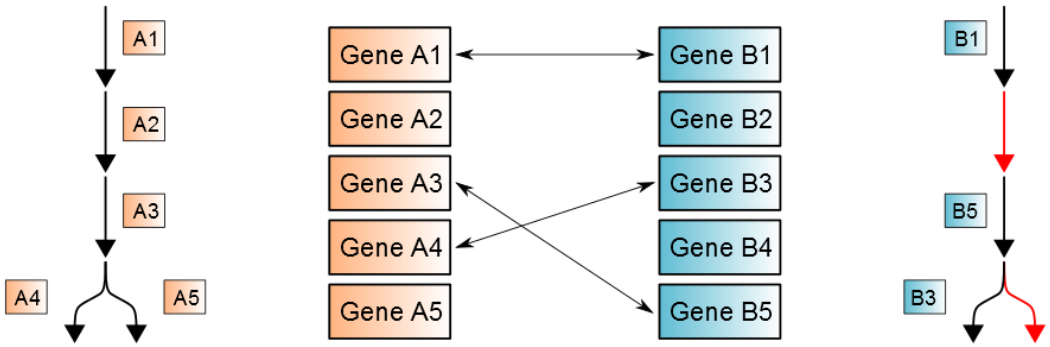

Figure 5.2: A visual representation of the proteome comparison between Organism A and Organism B. Genes in Organism A are matched with corresponding genes in Organism B, creating a mapping between the two genomes illustrated by the double-sided arrows. Using this mapping and the reaction network for Organism A (on left), the matching reads can be mapped onto the same network for Organism B (on right). Reactions with no genetic evidence in Organism B are shown in **red**, flagging them as prime candidates for removal and potentially creating gaps in the network.

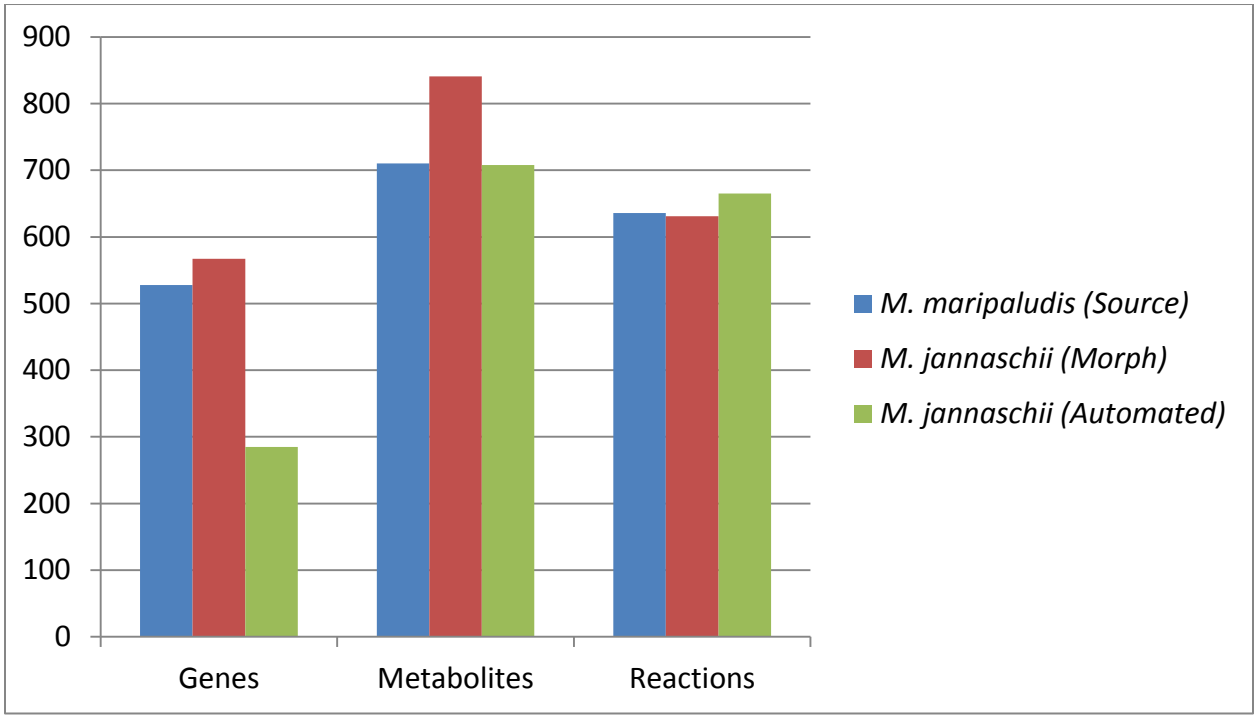

Figure 5.3: A comparison of basic model features for *M. jannaschii*. This includes the manual source model of *M. maripaludis*, our final morphed model of *M. jannaschii*, and an automated reconstruction of *M. jannaschii* generated using Kbase. The final morph contains more genes and many more metabolites than either of the other models, but slightly fewer reactions.

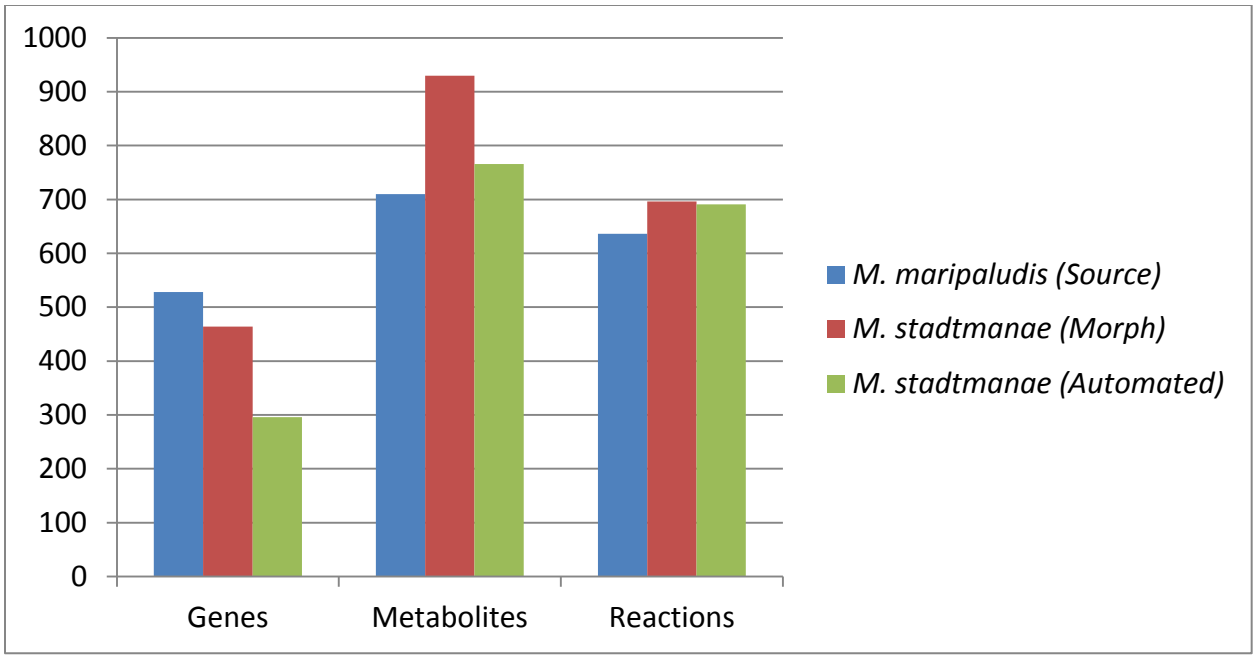

Figure 5.4: A comparison of basic model features for *M. stadtmanae*. This includes the manual source model of *M. maripaludis*, our final morphed model of *M. stadtmanae*, and an automated reconstruction of *M. stadtmanae* generated using Kbase. The final morph contains fewer genes than the source model but more than the automated model. It also contains many more metabolites and slightly more reactions than either of the other models.

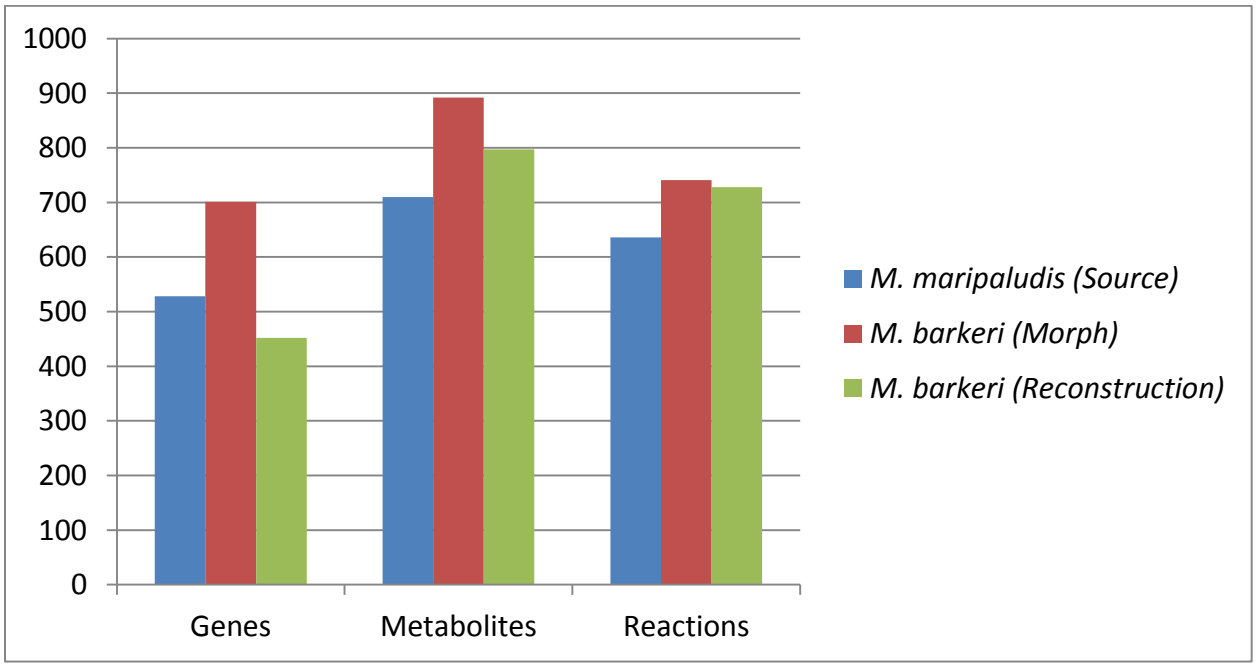

Figure 5.5: A comparison of basic model features for *M. barkeri*. This includes the manual source model of *M. maripaludis*, our final morphed model of *M. barkeri*, and an automated reconstruction of *M. barkeri* generated using Kbase. The final morph contains fewer genes than the source model but more than the automated model. It also contains many more metabolites and slightly more reactions than either of the other models.

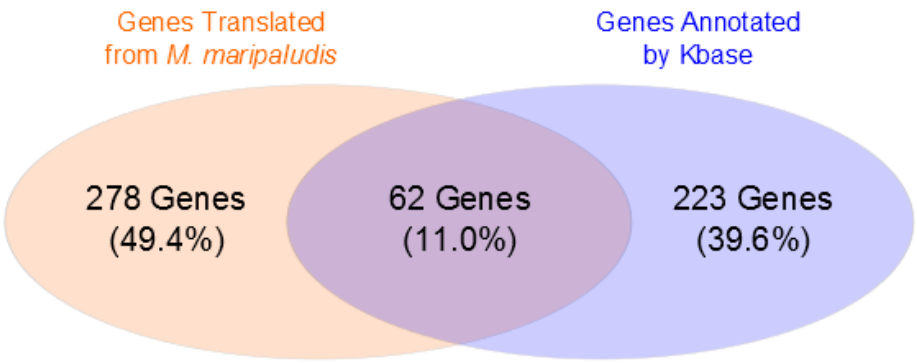

Figure 5.6: Gene origins for *M. jannaschii* morphed model. Percentages reflect the portion of total genes contributed by a particular subcategory (e.g. 11.0% of the genes in the *M. jannaschii* morph are from both source models)

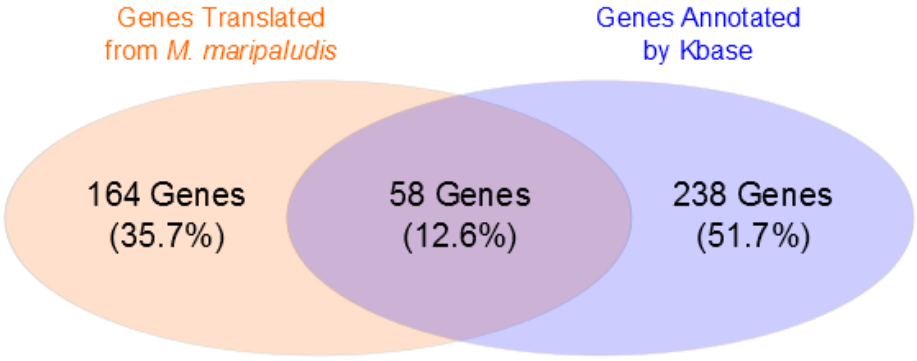

Figure 5.7: Gene origins for *M. stadtmanae* morphed model. Percentages reflect the portion of total genes contributed by a particular subcategory (e.g. 12.0% of the genes in the *M.stadtmanae* morph are from both source models)

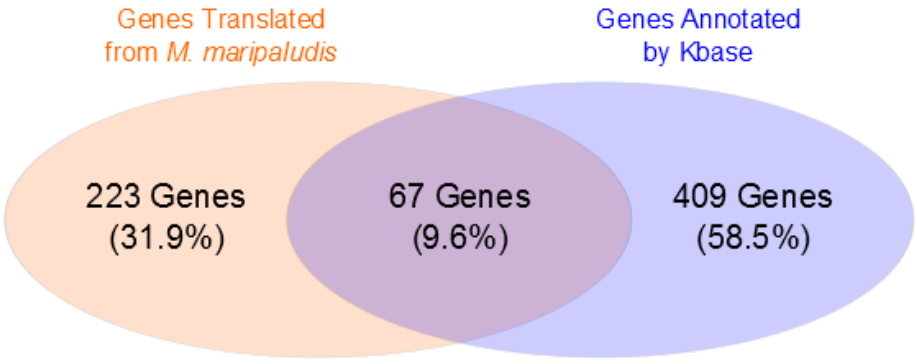

Figure 5.8: Gene origins for *M. barkeri* morphed model. Percentages reflect the portion of total genes contributed by a particular subcategory (e.g. 9.6% of the genes in the *M. barkeri* morph are from both source models)

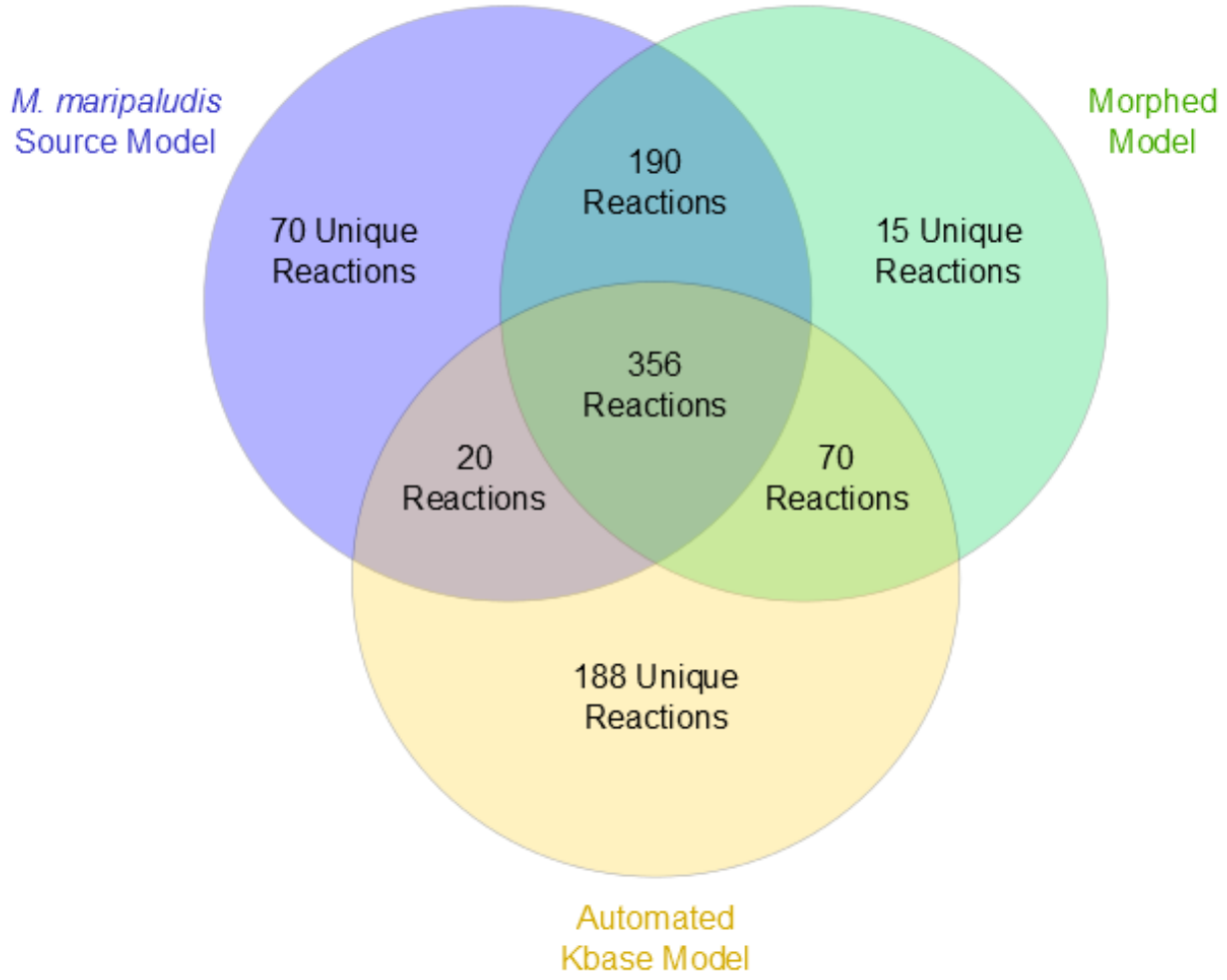

Figure 5.9: Reaction overlap for *M. jannaschii* model forms. Gene associations for reations in the final morph can be found in Table 5.2

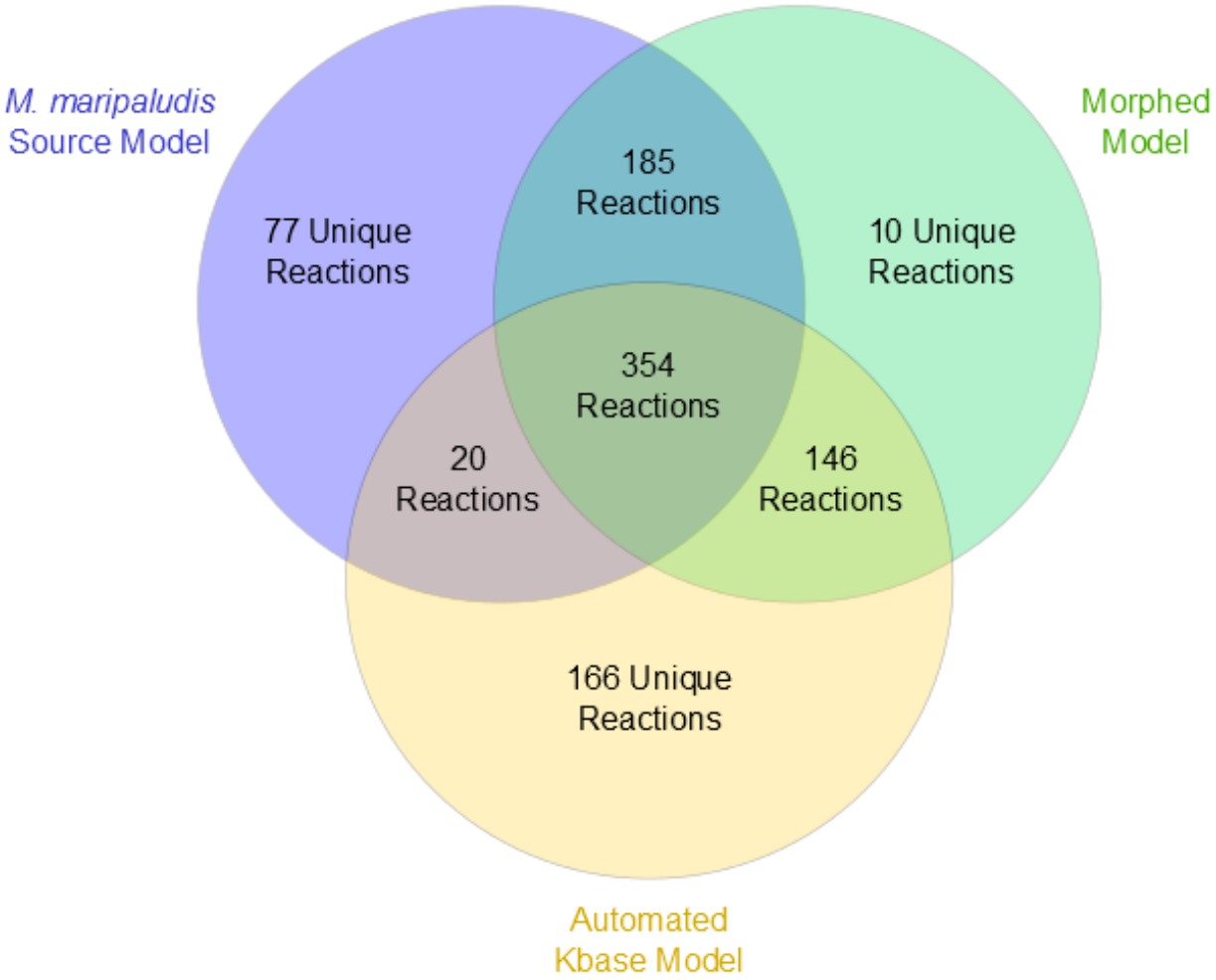

Figure 5.10: Reaction overlap for *M. stadtmanae* model forms. Gene associations for reations in the final morph can be found in Table 5.2

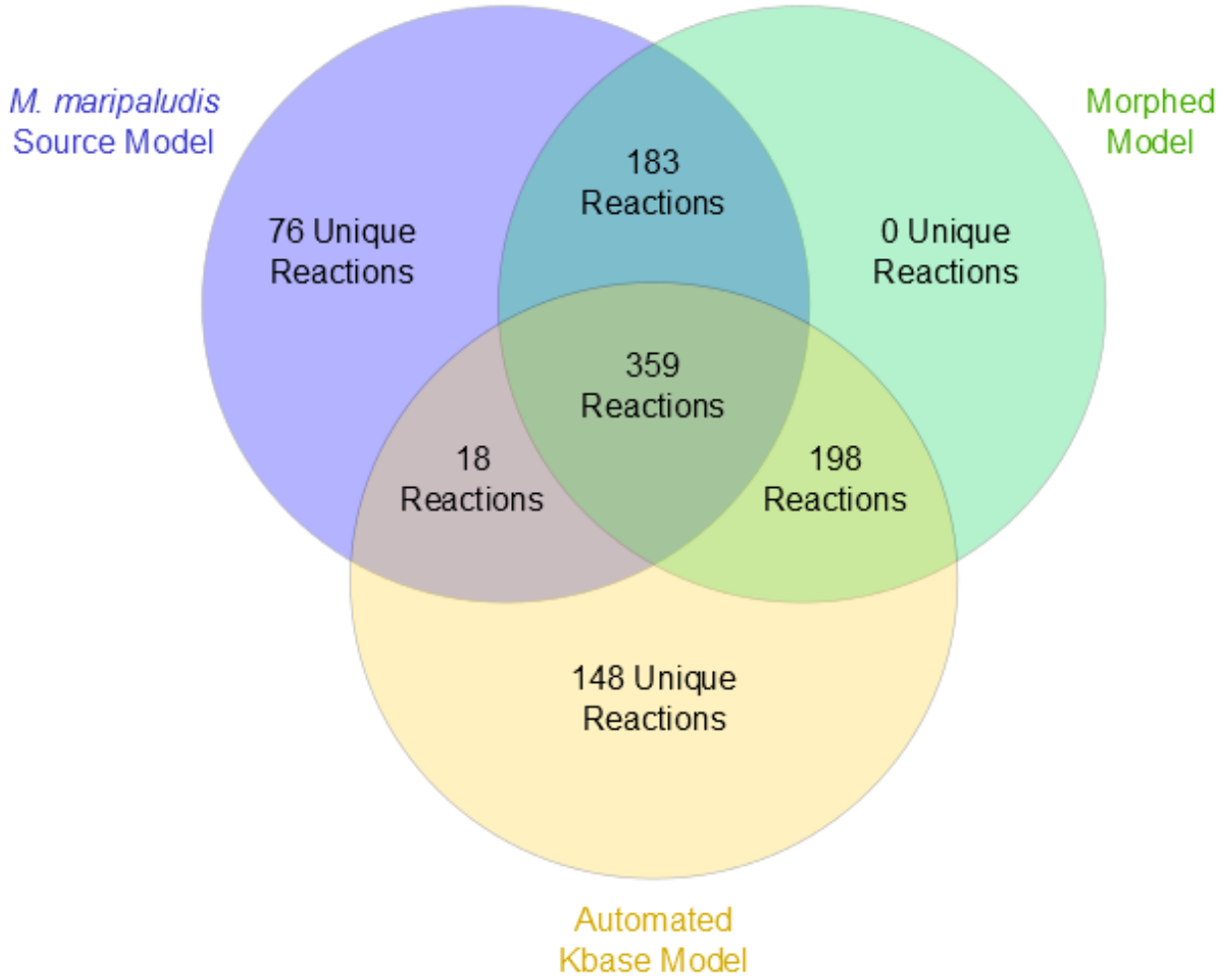

Figure 5.11: Reaction overlap for *M. barkeri* model forms. Gene associations for reations in the final morph can be found in Table 5.2

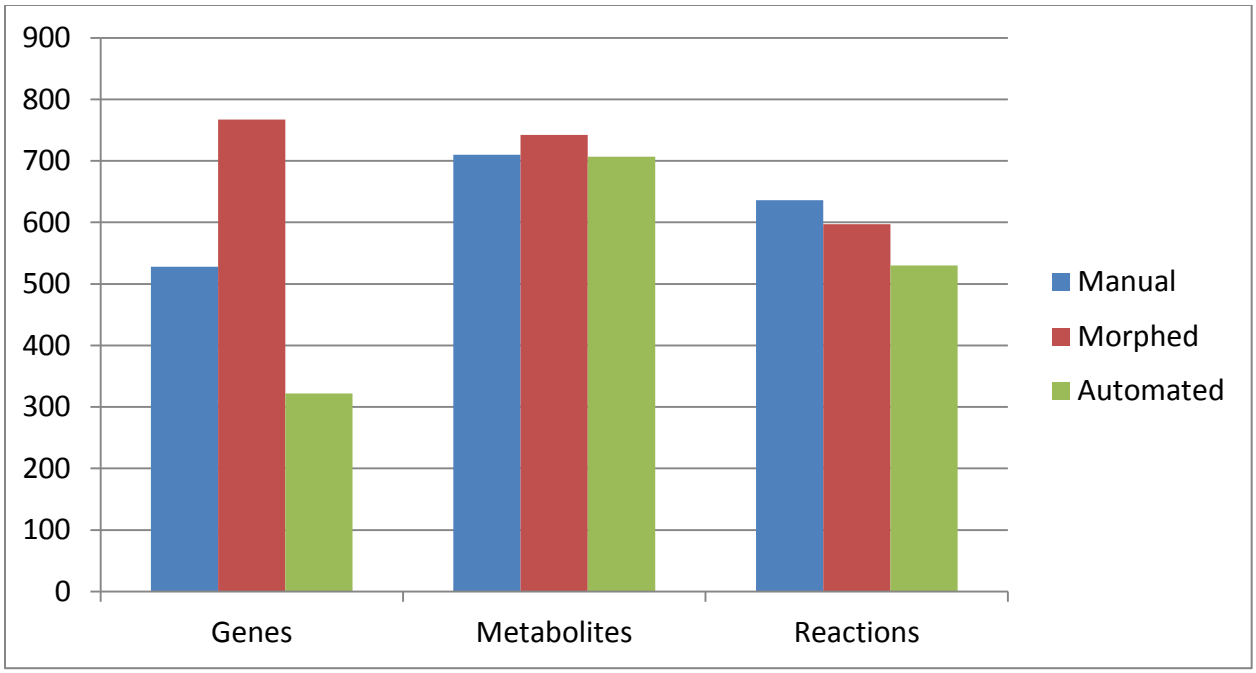

Figure 5.12: A comparison of basic model features for *M. maripaludis.* This includes the manual source model of *M. maripaludis*, our final morphed model of *M. stadtmanae*, and an automated reconstruction of *M. maripaludis* generated using Kbase. The final morph contains many more genes and slightly more metabolites than either of the other models. It contains fewer reactions than the manual model, but more than the automated model.

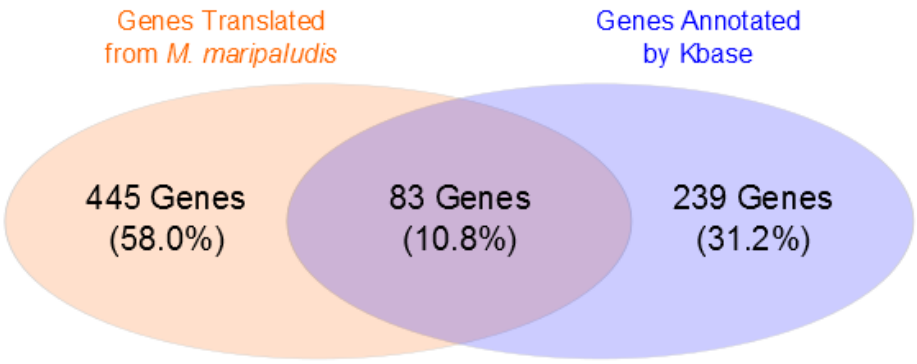

Figure 5.13: Gene origins for *M.maripaludis* morphed model. Percentages reflect the portion of total genes contributed by a particular subcategory (e.g. 10.8% of the genes in the *M. maripaludis* morph are from both source models)

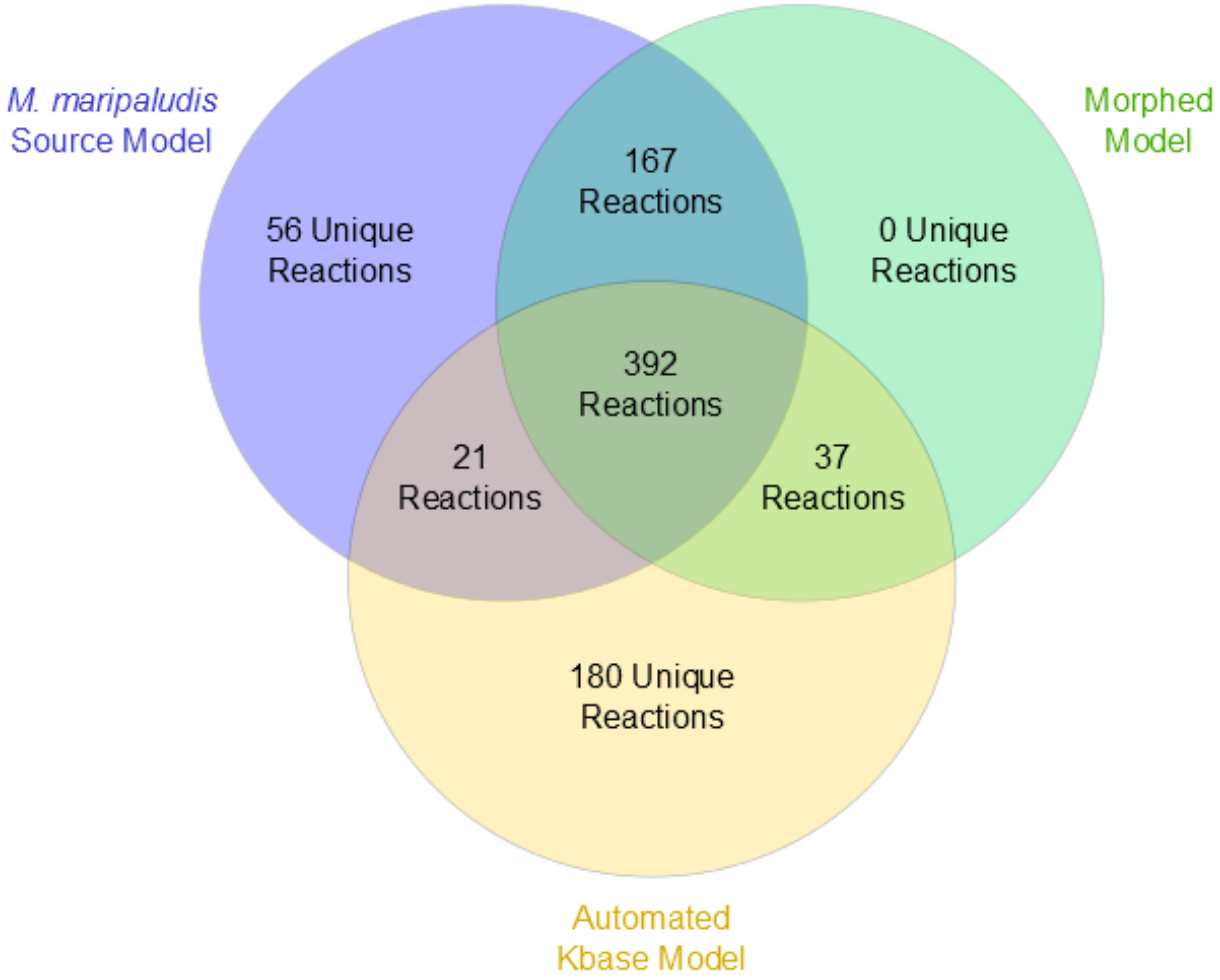

Figure 5.14: Reaction overlap for *M. maripaludis* model forms. Gene associations for reations in the final morph can be found in Table 5.2

Chapter 6: Discussion and Conclusions

In the previous chapters, I described various efforts to increase usage of manual data in concert with automated approaches. Though these projects all aimed to explore the need for developing more techniques that blend manual and automated methods, they are all threads in the much larger tapestry of improving metabolic engineering. From the standpoint of chemical engineering, biological systems present a unique challenge when compared to traditional chemical processes. Unlike purely chemical systems, which have generally been known for decades, living systems present a whole new litany of poorly understood phenomena. Thus as we try to engineer such systems to perform desired metabolic processes akin to those we can achieve chemically, we are encountering novel obstacles that require much more characterization.

My investigations into chemically defined media and metabolic reconstructions are both important pieces to improve our community's capabilities for metabolic engineering. Media formulations are the substrates for our proposed processes, the raw materials that ultimately dictate the possible products of our systems. Hence, it is essential that as we hone our metabolic engineering tools, we greatly expand upon our understanding of what organisms use to grow and why. If media conditions are the raw materials to a biological process, then metabolic reconstructions and models are a facsimile of the unit operations in a chemical process, the series of mechanisms through which the inputs reach their end products. Although the reconstruction is innately bound by available substrates in media formulations, the chemical reactions therein are the engine that drives overall transformations. The two are inherently linked to one another, necessitating a better understanding of both if we hope to truly master metabolic engineering.

The work I presented in the previous chapters underscored the importance of leveraging all available knowledge for formulating media and modeling metabolism. Summing up the knowledge I have gained from my studies, I believe there are several conclusions to be drawn regarding both my completed work. These takeaways have salient implications for the future of media development and metabolic modeling, as well as for metabolic engineering as a whole.

Takeaways from Completed Work

Balancing Databases and Other Resources

Most of the work I have presented relates in some way to using databases to organize and centralize information. As computational resources rapidly grow in their processing and storage capabilities, databases are increasingly thrust into a central role, both inside and outside of science. There is much to be said for a database's ability to bring together many disparate information sources, a benefit I harnessed when creating MediaDB. At the outset of that project, two things seemed readily apparent: (1) the biological community was acutely aware of the fact that the unknown microbial world greatly exceeds the known one; (2) details on how to grow the characterized part of the biosphere were scattered throughout published literature. Putting together those two observations, it was clear that a database could help accrue much of this information, addressing what I saw as a deficiency in how we store microbial growth media. MediaDB was the first step toward consolidating that information and making it more widely accessible to researchers without requiring them to scour countless reference materials. By bringing together many sources to create a larger dataset, my database enabled larger scale analyses of growth media that were previously more difficult due to the widely dispersed nature of media information.

Databases in the metabolic modeling space fill a similar function, gathering many different genome annotations into centralized resources that fuel high throughput analyses. And yet, my work with these

databases has served to illustrate many of the dangers involved with relying solely upon them. The differences between my manually curated model, iMR540, and a model of *M. maripaludis* created using automated reconstruction demonstrated how much information we could lose if we neglect manual curation, and how much clearly remains unknown about cellular metabolisms, even after extensive studies. That is not to say that there is no value in the automated reconstruction, which served as the basis for my manual model and contributed many genes that I could not have found through literature searches. But it is crucial to recognize the limitations of any database, particularly that the understanding gained from the database is bound by the information contained therein.

Creating my own model required adding reactions and genes not contained in the Department of Energy Knowledgebase (Kbase; www.kbase.us) to supplement the database information. Else, I would not have captured the same depth of biochemical information inherent in my finished model. Correspondingly, MediaDB could be a useful resource for studying trends in media formulation, perhaps for creating a new medium to culture a novel organism. However, using MediaDB alone would likely miss many details not present in the database and would have much lower likelihood of success than an approach that encompasses more sources. Databases are excellent at compiling data, but it is critical that database users not be blinded to external resources. When using such databases it is vital to recognize their limitations and seek out information from other resources. For example, one major limitation is that negative results from media that were tried unsuccessfully are rarely published, and thus do not often appear in databases. The key is finding a way to achieve a balance of resources, with database information providing a solid starting foundation and external literature sources filling in knowledge gaps to create a high quality end product.

Applying General Knowledge to Lesser Known Organisms

A readily apparent feature of biological databases is the degree to which they are dependent on a relatively small core group of organisms. We typically refer to these as “model organisms”, oftentimes industrial workhorses like *Escherichia coli* and *Saccharomyces cerevisiae*, but also representatives from other clades, such as the *Methanosarcina* genus for methanogens. If we critically examine the organisms in MediaDB for example, the 200+ strains we collected comprise a very small group of organisms, especially when compared to the nearly 58,000 organisms in RefSeq. My studies have highlighted some of the consequences involved with applying knowledge gleaned from these organisms to lesser known species.

In addition to containing media from a small portion of known organisms, MediaDB displays fairly low diversity in media compounds. Perhaps the largest takeaway from studying the data in MediaDB was that the compounds we saw most frequently in media were those that appear in trace mineral and vitamin solutions. Indeed, these types of solutions are rather ubiquitous in traditional growth media development and provide many of the same nutrients to every organism. Presumably, the aim of such solutions is to fill the organism’s every need by supplying it with every conceivable component. This has worked quite well for the organisms we have grown thus far, allowing us to culture numerous organisms without necessarily determining what in the multi-component solutions is actually required for growth. However, this subset of organisms that has been cultured is a necessarily biased subset, in part based on somewhat arbitrary and pervasive choices of similar growth media.

There are a couple of problems that arise out of this approach. First, because we have used many of the same compounds in different media, it is difficult to study trends across different organisms. When I attempted to cluster our organisms based on their media compounds, I was unable to discover any particular pattern, largely because many disparate organisms shared many of the same components.

This complicates our ability to infer new media formulations from existing ones, because known media may lack defining characteristics that could inform new media designs. Second, our success with growing many organisms in similar media could actually prove to be a hindrance when trying to culture new isolates. One reason we have been able to culture the subset of organisms we have isolated to this point is because most of them grow well on rich media with lots of available nutrients. This success has primed me to believe that culturing a new organism comes down to adding more components to an established medium; to try to add the missing pieces to a mostly finished puzzle. But it is entirely possible that reason behind our inability to culture most organisms is that they employ markedly different growth strategies than the organisms we have already isolated. If we continue to employ strategies that encourage growth of microbial "weeds", we will succeed in growing more weeds, but may very well fail at coaxing growth out of our actual targets. Thus, as I consider the information contained within MediaDB, I realize this information may not apply to novel systems and may only be descriptive of the organisms contained therein.

My experience with metabolic modeling echoes the same message drawn from assembling and analyzing MediaDB. Just as media formulations tend to contain many of the same compounds, metabolic models and reconstructions tend to contain many of the same reactions, a point that was well illustrated in a review by Monk *et al*. [23]. Despite the accelerated growth of completed metabolic reconstructions, the majority of known enzymatic reactions have never appeared in a published reconstruction. Undoubtedly, this is partially due to the relatively small scope of organisms that have been modeled to this point; the roughly 200 different modeled organisms pales in comparison to the nearly 58,000 complete genomes. Hence, organisms lacking models likely contain many reactions that we simply have not encountered to this point. However, I would argue that the larger factor is that most models are heavily based on existing models from model organisms. In the aforementioned review, the

authors show that most models fall very close to *E. coli* in terms of metabolites and reactions, demonstrating the extent to which other organisms merely borrow annotations from the *E. coli* model.

Even when expanding out to the more broad range of model organisms that form the basis for annotation databases like Kbase, we cover a fairly narrow scope of gene annotations. We illustrated this poignantly in Chapter 5 by showing how much our manual model of *Methanococcus maripaludis* diverged from the automated Kbase model. Using automated annotations from a database may work well for organisms that are similar to model organisms, but for something further from the beaten path, these same annotations appeared to fit much more poorly. This is not necessarily surprising; we would expect that although databases and automated approaches work well on new test cases that bear strong similarity to the training set, they can fair quite poorly on test cases that are quite different. Such methods are inherently biased by their training sets, thus when we apply a new test case, we get back information that looks like what we have encountered before. But as we look to expand our knowledge about biological processes, we must necessarily break free of relying on these methods, lest we lose out on the increased diversity we could add by using more meticulous manual methods.

Recognizing the Limitations of Predictive Automated Methods for Complex Problems

With the boom in sequencing information covered in Chapter 1, we now have more genomic data than ever and enormous potential for applying these data to address some quite challenging problems. Though this wealth of data presents numerous opportunities to discover new phenomena and emergent properties, we must be cautious when evaluating predictive models and methods that utilize these data. The sheer volume of analyses we can perform with such tools are valuable to be sure, but it would be folly to believe they can capture the full range of phenomena occurring within a biological system. Rather, automated methods relying on large datasets give some idea of a subset of these phenomena and rely on rigorous manual methods to fill in the rest.

It is somewhat sobering to look back at the origins of the MediaDB project and recall that my initial goal was to create a method that predicts growth media directly from sequenced genomes. Drawn in by the potential power of this proposed automated method, I did not necessarily consider the many different factors that can impact culturing. I assumed that if given a large enough training set, I could devise some model that would work toward my end. Looking back at this juncture, my intentions were somewhat naïve because they oversimplified the culturing problem, conjecturing that the only missing piece was a bridge between genome and media formulations. Media are influenced not only by substrates present, but also substrate concentrations, pH, temperature, signaling compounds, and a host of other factors. Having now considered the limitations of what we can predict based on our limited knowledge, such a method can offer a starting set of candidate growth substrates but should not be counted on to produce working media formulations. Designing working growth media still requires manual work, both to scour reference materials for factors other than substrate diversity and to troubleshoot candidate formulations in the experimental lab.

As for metabolic reconstructions, the bevy of functions served by a finished reconstruction does not include the ability to completely determine metabolic engineering strategies. If we closely examine metabolic modeling and its application to strain design, it is remarkable to note the relatively small number of success stories compared to the number of completed models and methods. That is not to say that such tools cannot serve useful functions; as I demonstrated in Chapter 4, I was able to use iMR540 to make a series of useful predictions for *M. maripaludis.* And yet, at the end of the day the results from those predictions can be somewhat unsatisfying because all of them still require verification from manual wet lab experiments. Our metabolic models may be powerful for contextualizing metabolic processes, but they overlook a wealth of biological phenomena to focus in on metabolism. A number of factors—protein expression levels, regulatory effects from environmental factors, unknown metabolic

pathways—could be responsible for our inability thus far to reproduce our predictions in actual cells and it is quite difficult to determine the factors responsible.

When combined with the database bias toward model organisms, the limited scope of automated methods cannot possibly predict the full spectrum of phenomena in uncharacterized organisms. Much as was the case when evaluating iMR540, it is important to temper our expectations of such methods and realistically understand their capabilities. Ultimately, these methods will not solve complex biological problems all on their own, but they can still serve as important cogs in more involved solutions. Metabolic models can help bound the space of metabolism, helping us predict what to expect from a system and giving us a rough idea of what is metabolically feasibly. Media prediction provides a starting point for designing new media that draws on successful formulations, cutting down on time spent scouring literature for these data. Each of these approaches can play an important role as we strive to better understand organisms for the purpose of metabolic engineering, so long as we do not rely solely on these methods.

Future Directions

The work I presented in this dissertation represents several steps toward untangling the web of metabolic phenomena in biological systems, yet much more remains to be done. If we hope to understand the mechanisms that drive metabolism well enough to intelligently design and optimize new strains, then it is crucial to continue to employ manual methods in concert with automated tools.

Much of this work must come in utilizing and expanding our existing tools. In the case of MediaDB, I have provided the first repository intended to compile defined growth media, but my database and others like it are only as useful as the information they contain. If we truly want to study existing media formulations and use that knowledge to formulate new media, we need a larger community buy-in to

media databases and high-throughput unbiased screens that present evenly both positive and negative growth outcomes. By contributing more data to these resources, we will increase our dataset, providing more chances to uncover emergent trends that govern organisms' nutrient requirements.

One aspect of growth media that often goes unmentioned is collecting negative results; that is, records of unsuccessful growth media. Biochemical literature is plentiful with examples of successful growth experiments and rarely contains details of failed culturing attempts. Yet these data could prove to be even more valuable than successful media formulations, providing known growth constraints in the form of toxic or inhibitory compounds. With the relatively recent introduction of Biolog phenotype microarrays for high-throughput culturing experiments, we now have an engine with which to quickly compile unsuccessful growth conditions. If we collect these data into a database similar to MediaDB, we can increase our chances of learning more about microbial growth requirements.

Similar to growth media experiments, the future of metabolic modeling for metabolic engineering is seemingly in the hands of the community. Judging by the perpetuation of recent perspective manuscripts about the state of metabolic modeling, modelers are becoming increasingly aware of the need to better organize ourselves. We have built many manually curated models without strictly enforcing standards on their formats and nomenclature, complicating comparative studies that use multiple models in disparate formats. As I laid out in Chapter 1, it is clear that we need to come together on a set of standards and a repository to house manually curated models. Doing so will increase the usefulness of future models and allow us to convert existing models into the same language. This development would better unify metabolic modelers, trivializing the process of using another group's model and lowering the barrier to using that model for comparative studies.

Arguably the greatest challenge to metabolic modeling is the aforementioned general nature of databases dominated by model organisms. When we base new models on gene annotations from these model organisms, it can sabotage our ability to use such models to discover new phenomena. We would do well to target a wider breadth of organisms that span the tree of life, particularly those that carry out unusual metabolic functions. If we put the effort into creating manually curated reconstructions for these organisms, particularly by relying heavily on biochemical characterization experiments, we can greatly enhance the richness in our annotation databases. Ultimately, such characterizations rely on wet lab experiments, necessitating that we continue to progress in our ability to isolate new organisms. Thus, diversifying metabolic models goes hand in hand with developing new growth media; we must progress substantially on both fronts if we hope to uncover more about the inner workings of microbial metabolism.

Coming back to metabolic engineering itself, I envision a future where we can design organisms completely from a computational platform. In this idealized future, I imagine we will compile sufficient metabolic information that for any desired set of substrates and products, we will be able query our database of metabolism to rank the best microbial host based on its growth rate, byproduct tolerance, and various other features. Though this sounds somewhat farfetched given the enormity of the unknown microbial world, we are working toward this goal every day. Each new growth medium we create, each manually curated metabolic network we build is a new piece in our understanding of microbial capabilities.

Based on the lessons taken from my dissertation work, it is clear that we still have a long way to go on this front. But by tapping into these lessons, I believe we can devise a better roadmap for reaching this goal than the path we are currently on. If we recognize the strengths and limitations of our data-rich automated tools and resources, we can better understand how to supplement these approaches with

manual data and open our eyes to the gaps in our knowledge rather than glossing over them. Employing such a strategy, where we utilize faster and more precise automated methods in concert with meticulous manual methods, we can move toward completely *in silico* strain design as a guiding force for metabolic engineering. Though we will likely never divorce ourselves completely from time-intensive manual methods, if we are cognizant of our abilities and deficiencies in computation, we can create hybrid methods that reap the benefits of both approaches and completely change the way we interact with the microbial world.

Chapter 7: References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. PNAS. 1977;74: 5463–5467.

2. Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG, Brown NL, et al. The nucleotide sequence of bacteriophage ϕX174. Journal of Molecular Biology. 1978;125: 225–246. doi:10.1016/0022-2836(78)90346-7

3. The NCBI Handbook. National Center for Biotechnology Information (US); 2002.

4. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature protocols. Nature Publishing Group; 2010;5: 93–121. doi:10.1038/nprot.2009.203

5. Simeonidis E, Price ND. Genome-scale modeling for metabolic engineering. Journal of industrial microbiology & biotechnology. 2015; 327–338. doi:10.1007/s10295-014-1576-3

6. Heavner BD, Smallbone K, Barker B, Mendes P, Walker LP. Yeast 5 – an expanded reconstruction of the Saccharomyces cerevisiae metabolic network. BMC Systems Biology. 2012;6: 55. doi:10.1186/1752-0509-6-55

7. Mardis ER. Next-Generation Sequencing Platforms. Annual Review of Analytical Chemistry. 2013;6: 287–303. doi:10.1146/annurev-anchem-062012-092628

8. Edwards JS, Palsson BO. Systems Properties of the *Haemophilus influenzae* Rd Metabolic Genotype. Journal of Biological Chemistry. 1999;274: 17410–17416.

9. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, et al. Fluorescence detection in automated DNA sequence analysis. Nature. 1986;321: 674–679. doi:10.1038/321674a0

10. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucl Acids Res. 2000;28: 27–30. doi:10.1093/nar/28.1.27

11. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucl Acids Res. 2004;32: D438–D442. doi:10.1093/nar/gkh100

12. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Academy of Sciences of the United States of America. National Academy of Sciences; 2007;104: 1777–1782.

13. Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet. 2010;11: 31–46. doi:10.1038/nrg2626

14. Marsh M, Tu O, Dolnik V, Roach D, Solomon N, Bechtol K, et al. High-throughput DNA sequencing on a capillary array electrophoresis system. J Capillary Electrophor. 1997;4: 83–89.

15. Satish Kumar V, Dasika MS, Maranas CD. Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics. 2007;8: 212. doi:10.1186/1471-2105-8-212

16. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotech. 2008;26: 1135–1145. doi:10.1038/nbt1486

17. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437: 376–380. doi:10.1038/nature03959

18. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucl Acids Res. 2006;34: e22–e22. doi:10.1093/nar/gnj023

19. Turcatti G, Romieu A, Fedurco M, Tairi A-P. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. Nucl Acids Res. 2008;36: e25–e25. doi:10.1093/nar/gkn021

20. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, et al. Single-Molecule DNA Sequencing of a Viral Genome. Science. 2008;320: 106–109. doi:10.1126/science.1150427

21. Braslavsky I, Hebert B, Kartalov E, Quake SR. Sequence information can be obtained from single DNA molecules. PNAS. 2003;100: 3960–3964. doi:10.1073/pnas.0230489100

22. Henry CS, DeJongh M, Best A a, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature biotechnology. Nature Publishing Group; 2010;28: 977–82. doi:10.1038/nbt.1672

23. Monk J, Nogales J, Palsson BO. Optimizing genome-scale network reconstructions. Nat Biotech. 2014;32: 447–452. doi:10.1038/nbt.2870

24. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010; ddq416. doi:10.1093/hmg/ddq416

25. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucl Acids Res. 2014;42: D459–D471. doi:10.1093/nar/gkt1103

26. Poolman MG, Bonde BK, Gevorgyan A, Patel HH, Fell DA. Challenges to be faced in the reconstruction of metabolic networks from public databases. Syst Biol (Stevenage). 2006;153: 379–384.

27. Milne CB, Eddy JA, Raju R, Ardekani S, Kim P-J, Senger RS, et al. Metabolic network reconstruction and genome-scale model of butanol-producing strain Clostridium beijerinckii NCIMB 8052. BMC Systems Biology. 2011;5: 130. doi:10.1186/1752-0509-5-130

28. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. PNAS. 1988;85: 2444–2448.

29. Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BFF. GenBank. Nucl Acids Res. 1998;26: 1–7. doi:10.1093/nar/26.1.1

30. Novère NL, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, et al. Minimum information requested in the annotation of biochemical models (MIRIAM). Nat Biotech. 2005;23: 1509–1515. doi:10.1038/nbt1156

31. Ravikrishnan A, Raman K. Critical assessment of genome-scale metabolic networks: the need for a unified standard. Brief Bioinform. 2015;16: 1057–1068. doi:10.1093/bib/bbv003

32. Chindelevitch L, Trigg J, Regev A, Berger B. An exact arithmetic toolbox for a consistent and reproducible structural analysis of metabolic network models. Nat Commun. 2014;5: 4893. doi:10.1038/ncomms5893

33. Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, et al. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nature biotechnology. 2008;26: 1155–60. doi:10.1038/nbt1492

34. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics. 2003;19: 524–531. doi:10.1093/bioinformatics/btg015

35. Olivier BG, Bergmann FT. SBML Level 3 Package: Flux Balance Constraints ('fbc') [Internet]. 11 Feb 2013 [cited 29 Oct 2015]. Available: http://resolver.caltech.edu/CaltechAUTHORS:20141028-172423140

36. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucl Acids Res. 2005;33: D501–D504. doi:10.1093/nar/gki025

37. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421. doi:10.1186/1471-2105-10-421

38. Li C, Donizelli M, Rodriguez N, Dharuri H, Endler L, Chelliah V, et al. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. BMC Systems Biology. 2010;4: 92. doi:10.1186/1752-0509-4-92

39. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucl Acids Res. 2015; gkv1049. doi:10.1093/nar/gkv1049

40. Heavner BD, Price ND. Transparency in metabolic network reconstruction enables scalable biological discovery. Current opinion in biotechnology. 2015;34C: 105–109. doi:10.1016/j.copbio.2014.12.010

41. Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, et al. SAR11 clade dominates ocean surface bacterioplankton communities. Nature. 2002;420: 806–810. doi:10.1038/nature01240

42. Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. Genetic diversity in Sargasso Sea bacterioplankton. Nature. 1990;345: 60–63. doi:10.1038/345060a0

43. Rappé MS, Connon SA, Vergin KL, Giovannoni SJ. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature. 2002;418: 630–633. doi:10.1038/nature00917

44. Carini P, Steindler L, Beszteri S, Giovannoni SJ. Nutrient requirements for growth of the extreme oligotroph "Candidatus Pelagibacter ubique" HTCC1062 on a defined medium. ISME J. 2013;7: 592–602. doi:10.1038/ismej.2012.122

45. Röling WFM, Ferrer M, Golyshin PN. Systems approaches to microbial communities and their functioning. Curr Opin Biotechnol. 2010;21: 532–538. doi:10.1016/j.copbio.2010.06.007

46. Richards MA, Cassen V, Heavner BD, Ajami NE, Herrmann A, Simeonidis E, et al. MediaDB: a database of microbial growth conditions in defined media. 2014; Available: http://dx.plos.org/10.1371/journal.pone.0103548

47. Oberhardt MA, Zarecki R, Gronow S, Lang E, Klenk H-P, Gophna U, et al. Harnessing the landscape of microbial culture media to predict new organism-media pairings. Nat Commun. 2015;6: 8493. doi:10.1038/ncomms9493

48. Borenstein E, Kupiec M, Feldman MW, Ruppin E. Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proc Natl Acad Sci U S A. 2008;105: 14482–14487. doi:10.1073/pnas.0806162105

49. Button DK, Schut F, Quang P, Martin R, Robertson BR. Viability and isolation of marine bacteria by dilution culture: theory, procedures, and initial results. Appl Environ Microbiol. 1993;59: 881–891.

50. D'Onofrio A, Crawford JM, Stewart EJ, Witt K, Gavrish E, Epstein S, et al. Siderophores from neighboring organisms promote the growth of uncultured bacteria. Chemistry & biology. Elsevier Ltd; 2010;17: 254–64. doi:10.1016/j.chembiol.2010.02.010

51. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotech. 2010;28: 977–982. doi:10.1038/nbt.1672

52. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protocols. 2010;5: 93–121. doi:10.1038/nprot.2009.203

53. Amann R. Who is out there? Microbial Aspects of Biodiversity. Systematic and Applied Microbiology. 2000;23: 1–8. doi:10.1016/S0723-2020(00)80039-9

54. Keller M, Zengler K. Tapping into microbial diversity. Nat Rev Micro. 2004;2: 141–150. doi:10.1038/nrmicro819

55. Alain K, Querellou J. Cultivating the uncultured: limits, advances and future challenges. Extremophiles. 2009;13: 583–594. doi:10.1007/s00792-009-0261-3

56. Vartoukian SR, Palmer RM, Wade WG. Strategies for culture of “unculturable” bacteria. FEMS Microbiol Lett. 2010;309: 1–7. doi:10.1111/j.1574-6968.2010.02000.x

57. Joint I, Mühling M, Querellou J. Culturing marine bacteria - an essential prerequisite for biodiscovery. Microb Biotechnol. 2010;3: 564–575. doi:10.1111/j.1751-7915.2010.00188.x

58. Pham VHT, Kim J. Cultivation of unculturable soil bacteria. Trends Biotechnol. 2012;30: 475–484. doi:10.1016/j.tibtech.2012.05.007

59. Prakash O, Shouche Y, Jangid K, Kostka JE. Microbial cultivation and the role of microbial resource centers in the omics era. Appl Microbiol Biotechnol. 2013;97: 51–62. doi:10.1007/s00253-012-4533-y

60. Kaeberlein T, Lewis K, Epstein SS. Isolating “uncultivable” microorganisms in pure culture in a simulated natural environment. Science. 2002;296: 1127–1129. doi:10.1126/science.1070633

61. Ferrari BC, Binnerup SJ, Gillings M. Microcolony cultivation on a soil substrate membrane system selects for previously uncultured soil bacteria. Appl Environ Microbiol. 2005;71: 8714–8720. doi:10.1128/AEM.71.12.8714-8720.2005

62. Yasumoto-Hirose M, Nishijima M, Ngirchechol MK, Kanoh K, Shizuri Y, Miki W. Isolation of marine bacteria by in situ culture on media-supplemented polyurethane foam. Mar Biotechnol. 2006;8: 227–237. doi:10.1007/s10126-005-5015-3

63. Bollmann A, Lewis K, Epstein SS. Incubation of environmental samples in a diffusion chamber increases the diversity of recovered isolates. Appl Environ Microbiol. 2007;73: 6386–6390. doi:10.1128/AEM.01309-07

64. Bruns A, Cypionka H, Overmann J. Cyclic AMP and acyl homoserine lactones increase the cultivation efficiency of heterotrophic bacteria from the central Baltic Sea. Appl Environ Microbiol. 2002;68: 3978–3987.

65. Bruns A, Nübel U, Cypionka H, Overmann J. Effect of signal compounds and incubation conditions on the culturability of freshwater bacterioplankton. Appl Environ Microbiol. 2003;69: 1980–1989.

66. Nichols D, Lewis K, Orjala J, Mo S, Ortenberg R, O’Connor P, et al. Short peptide induces an “uncultivable” microorganism to grow in vitro. Appl Environ Microbiol. 2008;74: 4889–4897. doi:10.1128/AEM.00393-08

67. D’Onofrio A, Crawford JM, Stewart EJ, Witt K, Gavrish E, Epstein S, et al. Siderophores from neighboring organisms promote the growth of uncultured bacteria. Chem Biol. 2010;17: 254–264. doi:10.1016/j.chembiol.2010.02.010

68. Janssen PH, Schuhmann A, Mörschel E, Rainey FA. Novel anaerobic ultramicrobacteria belonging to the Verrucomicrobiales lineage of bacterial descent isolated by dilution culture from anoxic rice paddy soil. Appl Environ Microbiol. 1997;63: 1382–1388.

69. Watve M, Shejval V, Sonawane C, Rahalkar M, Matapurkar A, Shouche Y, et al. The “K” selected oligophilic bacteria: a key to uncultured diversity? Current Science. 2000;78: 1535–1542.

70. Janssen PH, Yates PS, Grinton BE, Taylor PM, Sait M. Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions Acidobacteria, Actinobacteria, Proteobacteria, and Verrucomicrobia. Appl Environ Microbiol. 2002;68: 2391–2396.

71. Connon SA, Giovannoni SJ. High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. Appl Environ Microbiol. 2002;68: 3878–3885.

72. Sangwan P, Kovac S, Davis KER, Sait M, Janssen PH. Detection and cultivation of soil verrucomicrobia. Appl Environ Microbiol. 2005;71: 8402–8410. doi:10.1128/AEM.71.12.8402-8410.2005

73. Sait M, Hugenholtz P, Janssen PH. Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys. Environ Microbiol. 2002;4: 654–666.

74. Stevenson BS, Eichorst SA, Wertz JT, Schmidt TM, Breznak JA. New strategies for cultivation and detection of previously uncultured microbes. Appl Environ Microbiol. 2004;70: 4748–4755. doi:10.1128/AEM.70.8.4748-4755.2004

75. Davis KER, Joseph SJ, Janssen PH. Effects of growth medium, inoculum size, and incubation time on culturability and isolation of soil bacteria. Appl Environ Microbiol. 2005;71: 826–834. doi:10.1128/AEM.71.2.826-834.2005

76. Stott MB, Crowe MA, Mountain BW, Smirnova AV, Hou S, Alam M, et al. Isolation of novel bacteria, including a candidate division, from geothermal soils in New Zealand. Environ Microbiol. 2008;10: 2030–2041. doi:10.1111/j.1462-2920.2008.01621.x

77. Zengler K, Toledo G, Rappe M, Elkins J, Mathur EJ, Short JM, et al. Cultivating the uncultured. Proc Natl Acad Sci USA. 2002;99: 15681–15686. doi:10.1073/pnas.252630999

78. Zengler K, Walcher M, Clark G, Haller I, Toledo G, Holland T, et al. High-throughput cultivation of microorganisms using microcapsules. Meth Enzymol. 2005;397: 124–130. doi:10.1016/S0076-6879(05)97007-9

79. Ingham CJ, Sprenkels A, Bomer J, Molenaar D, van den Berg A, van Hylckama Vlieg JET, et al. The micro-Petri dish, a million-well growth chip for the culture and high-throughput screening of microorganisms. Proc Natl Acad Sci USA. 2007;104: 18217–18222. doi:10.1073/pnas.0701693104

80. Song H, Kim TY, Choi B-K, Choi SJ, Nielsen LK, Chang HN, et al. Development of chemically defined medium for Mannheimia succiniciproducens based on its genome sequence. Appl Microbiol Biotechnol. 2008;79: 263–272. doi:10.1007/s00253-008-1425-2

81. Price ND, Reed JL, Palsson BØ. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nature Reviews Microbiology. 2004;2: 886–897. doi:10.1038/nrmicro1023

82. Covert MW, Famili I, Palsson BO. Identifying constraints that govern cell behavior: a key to converting conceptual to computational models in biology? Biotechnol Bioeng. 2003;84: 763–772. doi:10.1002/bit.10849

83. Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. Current Opinion in Biotechnology. 2003;14: 491–496. doi:10.1016/j.copbio.2003.08.001

84. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ. Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol. 2009;7: 129–143. doi:10.1038/nrmicro1949

85. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucl Acids Res. 1999;27: 29–34. doi:10.1093/nar/27.1.29

86. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H-Y, Cohoon M, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005;33: 5691–5702. doi:10.1093/nar/gki866

87. Schellenberger J, Park JO, Conrad TM, Palsson BØ. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC bioinformatics. 2010;11: 213.

88. Hastings J, Matos P de, Dekker A, Ennis M, Harsha B, Kale N, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucl Acids Res. 2013;41: D456–D463. doi:10.1093/nar/gks1146

89. Bolton EE, Wang Y, Thiessen PA, Bryant SH. Chapter 12 - PubChem: Integrated Platform of Small Molecules and Biological Activities. In: Spellmeyer RAW and DC, editor. Annual Reports in Computational Chemistry. Elsevier; 2008. pp. 217–241. Available: http://www.sciencedirect.com/science/article/pii/S1574140008000121

90. Imanishi T, Nakaoka H. Hyperlink Management System and ID Converter System: enabling maintenance-free hyperlinks among major biological databases. Nucl Acids Res. 2009;37: W17–W22. doi:10.1093/nar/gkp355

91. Reed JL, Vo TD, Schilling CH, Palsson BO. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol. 2003;4: R54. doi:10.1186/gb-2003-4-9-r54

92. Benedict MN, Gonnerman MC, Metcalf WW, Price ND. Genome-Scale Metabolic Reconstruction and Hypothesis Testing in the Methanogenic Archaeon *Methanosarcina acetivorans* C2A. Journal of bacteriology. 2012;194: 855–865.

93. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41: D590–596. doi:10.1093/nar/gks1219

94. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22: 4673–4680.

95. Felsenstein J. {PHYLIP}: phylogenetic inference package, version 3.5c. Department of Genetics, University of Washington; 1993.

96. Subramaniam S. The Biology Workbench--a seamless database and analysis environment for the biologist. Proteins. 1998;32: 1–2.

97. Haynes CA, Gonzalez R. Rethinking biological activation of methane and conversion to liquid fuels. Nat Chem Biol. 2014;10: 331–339. doi:10.1038/nchembio.1509

98. Levi M. Climate consequences of natural gas as a bridge fuel. Climatic Change. 2013;118: 609–623. doi:10.1007/s10584-012-0658-3

99. Mueller TJ, Grisewood MJ, Nazem-Bokaee H, Gopalakrishnan S, Ferry JG, Wood TK, et al. Methane oxidation by anaerobic archaea for conversion to liquid fuels. J Ind Microbiol Biotechnol. 2014;42: 391–401. doi:10.1007/s10295-014-1548-7

100. Montzka SA, Dlugokencky EJ, Butler JH. Non-CO2 greenhouse gases and climate change. Nature. 2011;476: 43–50. doi:10.1038/nature10322

101. Kirschke S, Bousquet P, Ciais P, Saunois M, Canadell JG, Dlugokencky EJ, et al. Three decades of global methane sources and sinks. Nature Geoscience. 2013;6: 813–823. doi:10.1038/ngeo1955

102. Thauer RK, Kaster A-K, Seedorf H, Buckel W, Hedderich R. Methanogenic archaea: ecologically relevant differences in energy conservation. Nature Reviews Microbiology. 2008;6: 579–591. doi:10.1038/nrmicro1931

103. DiMarco AA, Bobik TA, Wolfe RS. Unusual coenzymes of methanogenesis. Annual review of biochemistry. 1990;59: 355–394.

104. structure of func of enzymes H2CO2 pathway 2002.pdf.

105. Costa KC, Leigh JA. Metabolic versatility in methanogens. Current Opinion in Biotechnology. 2014;29: 70–75. doi:10.1016/j.copbio.2014.02.012

106. Welte C, Deppenmeier U. Bioenergetics and anaerobic respiratory chains of aceticlastic methanogens. Biochimica et Biophysica Acta (BBA) - Bioenergetics. 2014;1837: 1130–1147. doi:10.1016/j.bbabio.2013.12.002

107. Heiden S, Hedderich R, Setzke E, Thauer RK. Purification of a cytochrome b containing H2:heterodisulfide oxidoreductase complex from membranes of *Methanosarcina barkeri*. European Journal of Biochemistry. 1993;213: 529–535. doi:10.1111/j.1432-1033.1993.tb17791.x

108. Kaster A-K, Moll J, Parey K, Thauer RK. Coupling of ferredoxin and heterodisulfide reduction via electron bifurcation in hydrogenotrophic methanogenic archaea. PNAS. 2011;108: 2981–2986. doi:10.1073/pnas.1016761108

109. Jones WJ, Paynter MJB, Gupta R. Characterization of *Methanococcus maripaludis* sp. nov., a new methanogen isolated from salt marsh sediment. Arch Microbiol. 1983;135: 91–97. doi:10.1007/BF00408015

110. Hendrickson EL, Kaul R, Zhou Y, Bovee D, Chapman P, Chung J, et al. Complete Genome Sequence of the Genetically Tractable Hydrogenotrophic Methanogen *Methanococcus maripaludis*. J Bacteriol. 2004;186: 6956–6969. doi:10.1128/JB.186.20.6956-6969.2004

111. Sarmiento FB, Leigh JA, Whitman WB. Chapter three - Genetic Systems for Hydrogenotrophic Methanogens. In: Ragsdale ACR and SW, editor. Methods in Enzymology. Academic Press; 2011. pp. 43–73. Available: http://www.sciencedirect.com/science/article/pii/B9780123851123000032

112. Graham DE, White RH. Elucidation of methanogenic coenzyme biosyntheses: from spectroscopy to genomics. Natural Product Reports. 2002;19: 133–147. doi:10.1039/b103714p

113. Stock T, Selzer M, Connery S, Seyhan D, Resch A, Rother M. Disruption and complementation of the selenocysteine biosynthesis pathway reveals a hierarchy of selenoprotein gene expression in the archaeon *Methanococcus maripaludis*. Molecular Microbiology. 2011;82: 734–747. doi:10.1111/j.1365-2958.2011.07850.x

114. Haydock AK, Porat I, Whitman WB, Leigh JA. Continuous culture of *Methanococcus maripaludis* under defined nutrient conditions. FEMS Microbiology Letters. 2004;238: 85–91. doi:10.1111/j.1574-6968.2004.tb09741.x

115. Hendrickson EL, Liu Y, Rosas-Sandoval G, Porat I, Soll D, Whitman WB, et al. Global Responses of Methanococcus maripaludis to Specific Nutrient Limitations and Growth Rate. Journal of Bacteriology. 2008;190: 2198–2205. doi:10.1128/JB.01805-07

116. Xia Q, Wang T, Hendrickson EL, Lie TJ, Hackett M, Leigh JA. Quantitative proteomics of nutrient limitation in the hydrogenotrophic methanogen Methanococcus maripaludis. BMC Microbiology. 2009;9: 149. doi:10.1186/1471-2180-9-149

117. Yoon SH, Turkarslan S, Reiss DJ, Pan M, Burn JA, Costa KC, et al. A systems level predictive model for global gene regulation of methanogenesis in a hydrogenotrophic methanogen. Genome Res. 2013;23: 1839–1851. doi:10.1101/gr.153916.112

118. Johnson EF, Mukhopadhyay B. Coenzyme F420-Dependent Sulfite Reductase-Enabled Sulfite Detoxification and Use of Sulfite as a Sole Sulfur Source by Methanococcus maripaludis. Applied and Environmental Microbiology. 2008;74: 3591–3595. doi:10.1128/AEM.00098-08

119. Lie TJ, Dodsworth JA, Nickle DC, Leigh JA. Diverse homologues of the archaeal repressor NrpR function similarly in nitrogen regulation. FEMS Microbiology Letters. 2007;271: 281–288. doi:10.1111/j.1574-6968.2007.00726.x

120. Simeonidis E, Price ND. Genome-scale modeling for metabolic engineering. J Ind Microbiol Biotechnol. 2015;42: 327–338. doi:10.1007/s10295-014-1576-3

121. Milne CB, Kim P-J, Eddy JA, Price ND. Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology. Biotechnology Journal. 2009;4: 1653–1670. doi:10.1002/biot.200900234

122. Stolyar S, Van Dien S, Hillesland KL, Pinel N, Lie TJ, Leigh JA, et al. Metabolic modeling of a mutualistic microbial community. Mol Syst Biol. 2007;3: 92. doi:10.1038/msb4100131

123. Goyal N, Widiastuti H, Karimi IA, Zhou Z. A genome-scale metabolic model of *Methanococcus maripaludis S2* for CO2 capture and conversion to methane. Mol Biosyst. 2014;10: 1043–1054. doi:10.1039/c3mb70421a

124. Susanti D, Mukhopadhyay B. An Intertwined Evolutionary History of Methanogenic Archaea and Sulfate Reduction. PLoS ONE. 2012;7: e45313. doi:10.1371/journal.pone.0045313

125. Graham DE, White RH. Elucidation of methanogenic coenzyme biosyntheses: from spectroscopy to genomics. Nat Prod Rep. 2002;19: 133–147. doi:10.1039/B103714P

126. Benedict MN, Mundy MB, Henry CS, Chia N, Price ND. Likelihood-Based Gene Annotations for Gap Filling and Quality Assessment in Genome-Scale Metabolic Models. PLoS Comput Biol. 2014;10: e1003882. doi:10.1371/journal.pcbi.1003882

127. Jackson BE, McInerney MJ. Anaerobic microbial metabolism can proceed close to thermodynamic limits. Nature. 2002;415: 454–456. doi:10.1038/415454a

128. Henry CS, Broadbelt LJ, Hatzimanikatis V. Thermodynamics-Based Metabolic Flux Analysis. Biophysical Journal. 2007;92: 1792–1805. doi:10.1529/biophysj.106.093138

129. Hoppe A, Hoffmann S, Holzhütter H-G. Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. BMC systems biology. 2007;1: 23.

130. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucl Acids Res. 2010;38: D473–D479. doi:10.1093/nar/gkp875

131. Feist AM, Palsson BO. The biomass objective function. Current Opinion in Microbiology. 2010;13: 344–349. doi:10.1016/j.mib.2010.03.003

132. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protocols. 2011;6: 1290–1307. doi:10.1038/nprot.2011.308

133. Heavner BD, Price ND. Transparency in metabolic network reconstruction enables scalable biological discovery. Current Opinion in Biotechnology. 2015;34: 105–109. doi:10.1016/j.copbio.2014.12.010

134. Kostromins A, Stalidzans E. Paint4Net: COBRA Toolbox extension for visualization of stoichiometric models of metabolism. Biosystems. 2012;109: 233–239. doi:10.1016/j.biosystems.2012.03.002

135. Porat I, Kim W, Hendrickson EL, Xia Q, Zhang Y, Wang T, et al. Disruption of the Operon Encoding Ehb Hydrogenase Limits Anabolic CO2 Assimilation in the Archaeon *Methanococcus maripaludis*. J Bacteriol. 2006;188: 1373–1380. doi:10.1128/JB.188.4.1373-1380.2006

136. Lie TJ, Costa KC, Lupa B, Korpole S, Whitman WB, Leigh JA. Essential anaplerotic role for the energy-converting hydrogenase Eha in hydrogenotrophic methanogenesis. PNAS. 2012;109: 15473–15478. doi:10.1073/pnas.1208779109

137. Lupa B, Hendrickson EL, Leigh JA, Whitman WB. Formate-Dependent H2 Production by the Mesophilic Methanogen *Methanococcus maripaludis*. Appl Environ Microbiol. 2008;74: 6584–6590. doi:10.1128/AEM.01455-08

138. Costa KC, Lie TJ, Jacobs MA, Leigh JA. H2-Independent Growth of the Hydrogenotrophic Methanogen *Methanococcus maripaludis*. mBio. 2013;4: e00062–13. doi:10.1128/mBio.00062-13

139. Costa KC, Wong PM, Wang T, Lie TJ, Dodsworth JA, Swanson I, et al. Protein complexing in a methanogen suggests electron bifurcation and electron delivery from formate to heterodisulfide reductase. PNAS. 2010;107: 11050–11055. doi:10.1073/pnas.1003653107

140. Hendrickson EL, Leigh JA. Roles of Coenzyme F420-Reducing Hydrogenases and Hydrogen- and F420-Dependent Methylenetetrahydromethanopterin Dehydrogenases in Reduction of F420 and Production of Hydrogen during Methanogenesis. J Bacteriol. 2008;190: 4818–4821. doi:10.1128/JB.00255-08

141. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. 1975;405: 442–451. doi:10.1016/0005-2795(75)90109-9

142. Flamholz A, Noor E, Bar-Even A, Milo R. eQuilibrator—the biochemical thermodynamics calculator. Nucl Acids Res. 2011; gkr874. doi:10.1093/nar/gkr874

143. Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V. Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks. Biophys J. 2008;95: 1487–1499. doi:10.1529/biophysj.107.124784

144. Costa KC, Yoon SH, Pan M, Burn JA, Baliga NS, Leigh JA. Effects of H2 and Formate on Growth Yield and Regulation of Methanogenesis in *Methanococcus maripaludis*. J Bacteriol. 2013;195: 1456–1462. doi:10.1128/JB.02141-12

145. Degtyarenko K, Matos P de, Ennis M, Hastings J, Zbinden M, McNaught A, et al. ChEBI: a database and ontology for chemical entities of biological interest. Nucl Acids Res. 2008;36: D344–D350. doi:10.1093/nar/gkm791

146. Setzke E, Hedderich R, Heiden S, Thauer RK. H2: heterodisulfide oxidoreductase complex from *Methanobacterium thermoautotrophicum*. European Journal of Biochemistry. 1994;220: 139–148. doi:10.1111/j.1432-1033.1994.tb18608.x

147. Thauer RK, Kaster A-K, Seedorf H, Buckel W, Hedderich R. Methanogenic archaea: ecologically relevant differences in energy conservation. Nature Reviews Microbiology. 2008;6: 579–591. doi:10.1038/nrmicro1931

148. Hedderich R, Thauer R k. Methanobacterium thermoautotrophicum contains a soluble enzyme system that specifically catalyzes the reduction of the heterodisulfide of coenzyme M and 7-mercaptoheptanoylthreonine phosphate with H2. FEBS Letters. 1988;234: 223–227. doi:10.1016/0014-5793(88)81339-5

149. Nitschke W, Russell MJ. Redox bifurcations: Mechanisms and importance to life now, and at its origin: A widespread means of energy conversion in biology unfolds…. BioEssays. 2012;34: 106–109. doi:10.1002/bies.201100134

150. Herrmann G, Jayamani E, Mai G, Buckel W. Energy Conservation via Electron-Transferring Flavoprotein in Anaerobic Bacteria. Journal of Bacteriology. 2008;190: 784–791. doi:10.1128/JB.01422-07

151. Costa KC, Wong PM, Wang T, Lie TJ, Dodsworth JA, Swanson I, et al. Protein complexing in a methanogen suggests electron bifurcation and electron delivery from formate to heterodisulfide reductase. PNAS. 2010;107: 11050–11055. doi:10.1073/pnas.1003653107

152. Thauer RK. The Wolfe cycle comes full circle. PNAS. 2012;109: 15084–15085. doi:10.1073/pnas.1213193109

153. Shieh JS, Whitman WB. Pathway of acetate assimilation in autotrophic and heterotrophic methanococci. J Bacteriol. 1987;169: 5327–5329.

154. Welander PV, Metcalf WW. Loss of the *mtr* operon in *Methanosarcina* blocks growth on methanol, but not methanogenesis, and reveals an unknown methanogenic pathway. Proceedings of the National Academy of Sciences of the United States of America. 2005;102: 10664–10669.

155. Kaster A-K, Goenrich M, Seedorf H, Liesegang H, Wollherr A, Gottschalk G, et al. More Than 200 Genes Required for Methane Formation from H2 and CO2 and Energy Conservation Are Present in *Methanothermobacter marburgensis* and *Methanothermobacter thermautotrophicus*. Archaea. 2011;2011: 1–23. doi:10.1155/2011/973848

156. DiMarco AA, Bobik TA, Wolfe RS. Unusual coenzymes of methanogenesis. Annual review of biochemistry. 1990;59: 355–394.

157. Siu S, Robotham A, Logan SM, Kelly JF, Uchida K, Aizawa S-I, et al. Evidence that Biosynthesis of the Second and Third Sugars of the Archaellin Tetrasaccharide in the Archaeon *Methanococcus maripaludis* Occurs by the Same Pathway Used by *Pseudomonas aeruginosa* To Make a Di-N-Acetylated Sugar. Metcalf WW, editor. Journal of Bacteriology. 2015;197: 1668–1680. doi:10.1128/JB.00040-15

158. Jain S, Caforio A, Driessen AJM. Biosynthesis of archaeal membrane ether lipids. Front Microbiol. 2014;5. doi:10.3389/fmicb.2014.00641

159. Balderston WL, Payne WJ. Inhibition of methanogenesis in salt marsh sediments and whole-cell suspensions of methanogenic bacteria by nitrogen oxides. Appl Environ Microbiol. 1976;32: 264–269.

160. Gonnerman MC, Benedict MN, Feist AM, Metcalf WW, Price ND. Genomically and biochemically accurate metabolic reconstruction of *Methanosarcina barkeri* Fusaro, iMG746. Biotechnology Journal. 2013;8: 1070–1079. doi:10.1002/biot.201200266

161. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, et al. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Molecular Systems Biology. 2007;3. doi:10.1038/msb4100155

162. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, et al. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism--2011. Molecular Systems Biology. 2014;7: 535–535. doi:10.1038/msb.2011.65

163. Sarmiento F, Mrázek J, Whitman WB. Genome-scale analysis of gene function in the hydrogenotrophic methanogenic archaeon *Methanococcus maripaludis*. PNAS. 2013;110: 4726–4731. doi:10.1073/pnas.1220225110

164. Unruh D, Pabst K, Schaub G. Fischer−Tropsch Synfuels from Biomass: Maximizing Carbon Efficiency and Hydrocarbon Yield. Energy Fuels. 2010;24: 2634–2641. doi:10.1021/ef9009185

165. Nauhaus K, Albrecht M, Elvert M, Boetius A, Widdel F. In vitro cell growth of marine archaeal-bacterial consortia during anaerobic oxidation of methane with sulfate. Environmental Microbiology. 2007;9: 187–196. doi:10.1111/j.1462-2920.2006.01127.x

166. Knittel K, Boetius A. Anaerobic Oxidation of Methane: Progress with an Unknown Process. Annual Review of Microbiology. 2009;63: 311–334. doi:10.1146/annurev.micro.61.080706.093130

167. Meulepas RJW, Jagersma CG, Gieteling J, Buisman CJN, Stams AJM, Lens PNL. Enrichment of anaerobic methanotrophs in sulfate-reducing membrane bioreactors. Biotechnol Bioeng. 2009;104: 458–470. doi:10.1002/bit.22412

168. Moran JJ, House CH, Freeman KH, Ferry JG. Trace methane oxidation studied in several Euryarchaeota under diverse conditions. Archaea. 2005;1: 303–309. doi:10.1155/2005/650670

169. Haroon MF, Hu S, Shi Y, Imelfort M, Keller J, Hugenholtz P, et al. Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage. Nature. 2013;500: 567–570. doi:10.1038/nature12375

170. Beal EJ, House CH, Orphan VJ. Manganese- and Iron-Dependent Marine Methane Oxidation. Science. 2009;325: 184–187. doi:10.1126/science.1169984

171. McGlynn SE, Chadwick GL, Kempes CP, Orphan VJ. Single cell activity reveals direct electron transfer in methanotrophic consortia. Nature. 2015;526: 531–535. doi:10.1038/nature15512

172. Wegener G, Krukenberg V, Riedel D, Tegetmeyer HE, Boetius A. Intercellular wiring enables electron transfer between methanotrophic archaea and bacteria. Nature. 2015;526: 587–590. doi:10.1038/nature15733

173. Elvert M, Suess E, Greinert J, Whiticar MJ. Archaea mediating anaerobic methane oxidation in deep-sea sediments at cold seeps of the eastern Aleutian subduction zone. Organic Geochemistry. 2000;31: 1175–1187. doi:10.1016/S0146-6380(00)00111-X

174. Kohler PRA, Metcalf WW. Genetic manipulation of Methanosarcina spp. Front Microbiol. 2012;3. doi:10.3389/fmicb.2012.00259

175. Pritchett MA, Metcalf WW. Genetic, physiological and biochemical characterization of multiple methanol methyltransferase isozymes in Methanosarcina acetivorans C2A. Molecular Microbiology. 2005;56: 1183–1194. doi:10.1111/j.1365-2958.2005.04616.x

176. Galagan JE, Nusbaum C, Roy A, Endrizzi MG, Macdonald P, FitzHugh W, et al. The Genome of M. acetivorans Reveals Extensive Metabolic and Physiological Diversity. Genome Res. 2002;12: 532–542. doi:10.1101/gr.223902

177. Moore BC, Leigh JA. Markerless Mutagenesis in *Methanococcus maripaludis* Demonstrates Roles for Alanine Dehydrogenase, Alanine Racemase, and Alanine Permease. J Bacteriol. 2005;187: 972–979. doi:10.1128/JB.187.3.972-979.2005

178. Walters AD, Smith SE, Chong JPJ. Shuttle Vector System for Methanococcus maripaludis with Improved Transformation Efficiency. Appl Environ Microbiol. 2011;77: 2549–2551. doi:10.1128/AEM.02919-10

179. King ZA, Feist AM. Optimizing Cofactor Specificity of Oxidoreductase Enzymes for the Generation of Microbial Production Strains—OptSwap. Industrial Biotechnology. 2013;9: 236–246. doi:10.1089/ind.2013.0005

180. Ghosh A, Zhao H, Price ND. Genome-scale consequences of cofactor balancing in engineered pentose utilization pathways in *Saccharomyces cerevisiae*. PLoS One. 2011;6: e27316.

181. Soo VWC, McAnulty MJ, Tripathi A, Zhu F, Zhang L, Hatzakis E, et al. Reversing methanogenesis to capture methane for liquid biofuel precursors. Microbial Cell Factories. 2016;15: 11. doi:10.1186/s12934-015-0397-z

182. Nazem-Bokaee H, Gopalakrishnan S, Ferry JG, Wood TK, Maranas CD. Assessing methanotrophy and carbon fixation for biofuel production by Methanosarcina acetivorans. Microbial Cell Factories. 2016;15: 10. doi:10.1186/s12934-015-0404-4

183. Oberhardt MA, Palsson BØ, Papin JA. Applications of genome-scale metabolic reconstructions. Molecular systems biology. 2009;5. Available: http://www.nature.com/msb/journal/v5/n1/full/msb200977.html

184. Zengler K, Palsson BO. A road map for the development of community systems (CoSy) biology. Nat Rev Micro. 2012;10: 366–372. doi:10.1038/nrmicro2763

185. Edwards JS, Palsson BO. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci USA. 2000;97: 5528–5533.

186. Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, et al. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat Biotech. 2008;26: 1155–1160. doi:10.1038/nbt1492

187. Heavner BD, Price ND. Comparative Analysis of Yeast Metabolic Network Models Highlights Progress, Opportunities for Metabolic Reconstruction. PLoS Comput Biol. 2015;11: e1004530. doi:10.1371/journal.pcbi.1004530

188. Oberhardt MA, Puchałka J, Martins dos Santos VAP, Papin JA. Reconciliation of Genome-Scale Metabolic Reconstructions for Comparative Systems Analysis. PLoS Comput Biol. 2011;7: e1001116. doi:10.1371/journal.pcbi.1001116

189. Benedict MN, Henriksen JR, Metcalf WW, Whitaker RJ, Price ND. ITEP: An integrated toolkit for exploration of microbial pan-genomes. BMC Genomics. 2014;15: 8. doi:10.1186/1471-2164-15-8

190. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics. 2008;9: 75. doi:10.1186/1471-2164-9-75

191. Kellis M, Patterson N, Birren B, Berger B, Lander ES. Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. J Comput Biol. 2004;11: 319–355. doi:10.1089/1066527041410319

192. Garcia J-L, Patel BK, Ollivier B. Taxonomic, phylogenetic, and ecological diversity of methanogenic Archaea. Anaerobe. 2000;6: 205–226.

Appendix A: Supporting Information for Chapter 2

Figure A.1: Full MediaDB schema. Dashed lines indicate foreign key relationships, oriented such that arrows point towards the referenced primary key. Each table is represented by a box headed by the table name and described by a list of column names and column types. This diagram was created using MySQL Workbench (www.mysql.com/products/workbench).

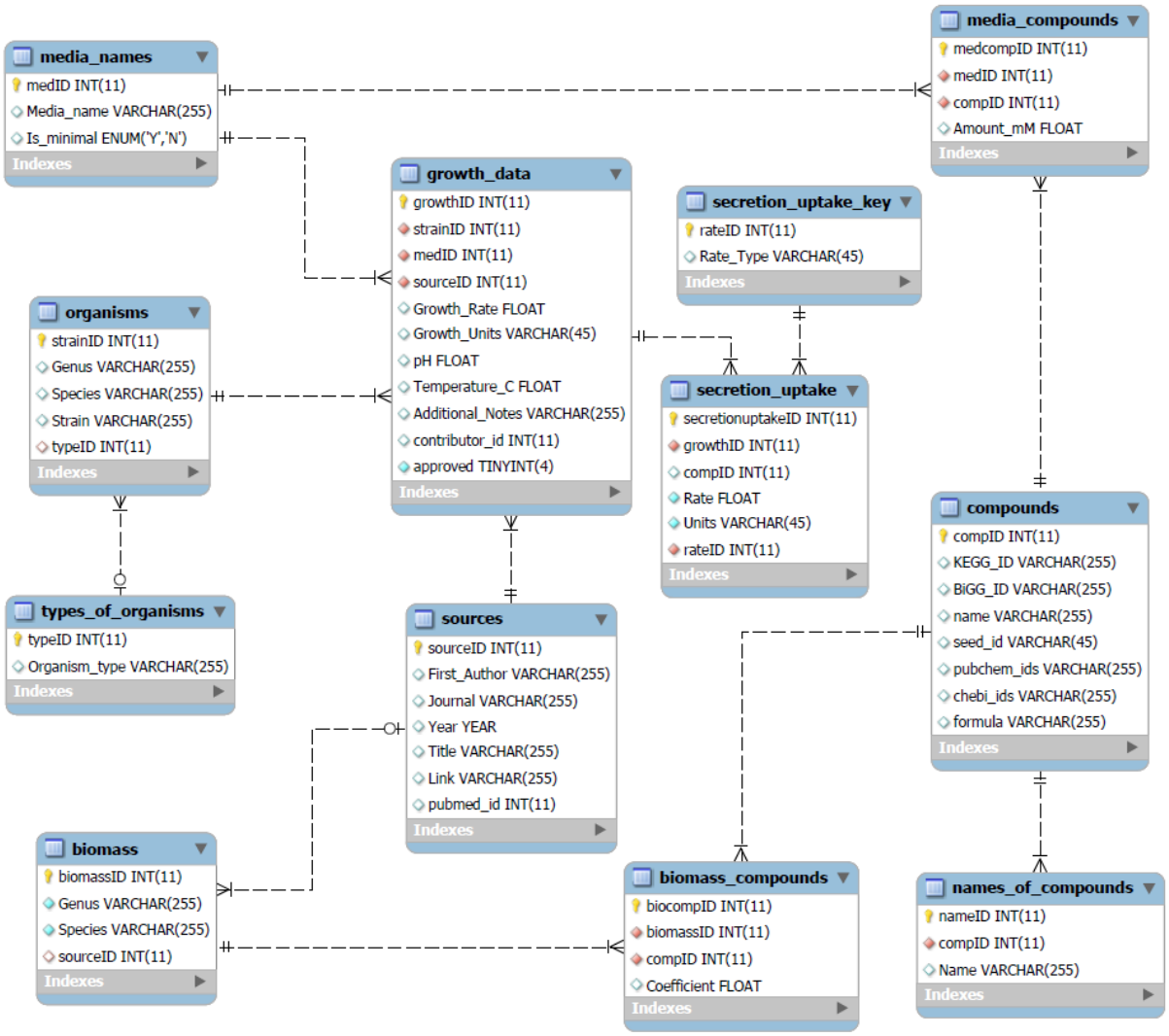

Table A.1: Full compound frequency results. Listed below is every compound that appears in at least one media formulation in MediaDB and the number of organism species known to utilize that compound (frequency).

Compound	Frequency
Calcium chloride anhydrous	49
Magnesium sulfate	41
Potassium dihydrogen phosphate	37
Sodium chloride	37
Ferrous sulfate	35
Zinc sulfate	35
Potassium dibasic phosphate	33
Ammonium chloride	30
Biotin	29
Cupric sulfate	28
Glucose	27
Sodium molybdate	24
Manganese sulfate	23
Boric acid	22
Magnesium chloride	22
Potassium chloride	22
Cobalt chloride	21
Cysteine	21
Manganese chloride	21
Thiamine HCl	20
Acetate	19
Arginine	19
Aspartate	19
4-Aminobenzoate	18
Glutamate	18
Leucine	18
Isoleucine	17
Histidine	16
Nicotinate	16
Phenylalanine	16
Alanine	15
Ammonium sulfate	15
Calcium pantothenate	15
Glycine	15
Lysine	15
Proline	15
Valine	15
Iron(III) chloride	14
Riboflavin	14

Table A.1 (cont.)

Compound	Frequency
Thiamine	14
Threonine	14
Folate	13
Cupric chloride	12
Dibasic sodium phosphate	12
Methionine	12
Vitamin B12	12
Citrate	11
Glycerol	11
Pyridoxine HCl	11
Asparagine	10
Nickel chloride	10
Zinc Chloride	10
Serine	9
Sodium sulfate	9
Tromethamine	9
Tryptophan	9
Tyrosine	9
Glutamine	8
Lipoate	8
Molybdic acid ammonium salt tetrahydrate	8
Nitrilotriacetic acid	8
Pantothenate	8
Sodium bicarbonate	8
Sodium citrate	8
DL-Serine	7
DL-Tyrosine	7
Pyruvate	7
Sodium dihydrogen phosphate	7
Urea	7
Cobaltous sulfate	6
EDTA	6
MOPS	6
Sodium nitrate	6
Uracil	6
Aluminum potassium sulfate	5
Ammonium nitrate	5
Ammonium phosphate	5
DL-methionine	5

Table A.1 (cont.)

Compound	Frequency
Ethanol	5
Fe(III)dicitrate	5
Potassium hydroxide	5
Pyridoxine	5
Resazurin	5
Succinate	5
Ascorbate	4
D-Fructose	4
D-Gluconic acid	4
D-Mannose	4
Inosine	4
Lactose	4
Potassium sulfate	4
Sodium borate	4
Sodium selenite	4
Sodium sulfide	4
Thymine	4
Tween 80	4
Xanthine	4
myo-Inositol	4
(S)-Malate	3
Adenine	3
Ammonium citrate	3
Cellobiose	3
Cl-	3
Ferrous chloride	3
Guanine	3
HEPES	3
Hemin	3
Lactate	3
Maltose	3
Methanol	3
Nitrate	3
Orotate	3
Pyridoxal HCl	3
Pyridoxamine HCl	3
Sodium EDTA	3
Sodium ammonium phosphate	3

Table A.1 (cont.)

Compound	Frequency
Sodium selenate	3
Sodium tungstate	3
Sucrose	3
Sulfate	3
Thymidine	3
(9Z)-Octadecenoic acid	2
(R)-Lactate	2
(S)-Lactate	2
1-Butanol	2
2-Methylpropanoate	2
3-Methylbutanoic acid	2
4-Hydroxy-L-proline	2
4-Hydroxybenzoate	2
ATP	2
Ammonium Acetate	2
Cobalt nitrate	2
D-Glutamate	2
Fe2+	2
Ferrous ammonium sulfate	2
Glutathione	2
Hypoxanthine	2
L-Sorbose	2
Magnesium	2
Mannitol	2
Menadione	2
NH4+	2
Nicotinamide	2
Orthophosphate	2
Pentanoate	2
Potassium	2
Potassium iodide	2
Potassium nitrate	2
Pyridoxamine	2
Sodium	2
Sodium Glutamate	2
Sodium L-lactate	2
Sodium Succinate	2
Sodium carbonate	2

Table A.1 (cont.)

Compound	Frequency
Sodium hydroxide	2
Sodium iodide	2
Sodium pyruvate	2
Starch	2
Toluene	2
Tricine	2
Xylose	2
Zinc Acetate	2
sodium silicate	2
(R,R)-Butane-2,3-diol	1
2-Mercaptoethanesulfonate	1
2-methylbutanoic acid	1
3,4-Dihydroxybenzoate	1
3,4-Dihydroxyphenylacetate	1
3-Hydroxypropanal	1
3-Sulfolactate	1
4-Cresol	1
4-Hydroxybenzaldehyde	1
4-Hydroxyphenylacetate	1
5-Hydroxyectoine	1
ACES	1
Adenosine sulfate	1
Ampicillin	1
Benzaldehyde	1
Benzoate	1
Benzyl alcohol	1
Borate	1
Butanoic acid	1
Calcium	1
Calcium carbonate	1
Calcium sulfate	1
Chloramphenicol	1
Cholesterol	1
Choline	1
Chromium(III) Chloride	1
Cobalt ion	1
Copper	1
Cyanocob(III)alamin	1

Table A.1 (cont.)

Compound	Frequency
D-Alanine	1
D-Aspartate	1
D-Galactose	1
D-Glucarate	1
D-Methionine	1
D-Ribose	1
D-Xylose	1
Deoxyribose	1
Diammonium tartrate	1
Diuron	1
EDDHA	1
Ectoine	1
Ethyl octanoate	1
Ethylene glycol	1
Fe3+	1
Ferric ammonium citrate	1
Ferric nitrate	1
Ferric nitrilotriacetate	1
Ferric oxide	1
Fumarate	1
GDP	1
Galactitol	1
Glycolate	1
Guanine HCl	1
Hexadecane	1
Hydrochloric acid	1
IPTG	1
Kanamycin	1
L-Arabinose	1
L-Citrulline	1
L-Inositol	1
L-Rhamnose	1
Maleic acid	1
Manganese	1
Manganese(IV) oxide	1
Molybdenum	1
Molybdenum trioxide	1
NAD+	1

Table A.1 (cont.)

Compound	Frequency
Nickel Sulfate	1
Nickel(II) ammonium sulfate	1
PIPES	1
Phenol	1
Polyvinyl alcohol	1
Potassium Gluconate	1
Propane-1-ol	1
Propanoate	1
Propenoate	1
Quinate	1
Retinol	1
Sodium Pantothenate	1
Sodium bromide	1
Sodium thiosulfate	1
Spectinomycin	1
Spermine phosphate	1
Streptomycin	1
Strontium chloride	1
TES	1
Triethanolamine	1
Uridine	1
Vitamin D3	1
Zinc	1
alpha,alpha-Trehalose	1
alpha-Tocopherol	1
beta-D-Fructose	1
dTMP	1
p-Hydroxybenzyl alcohol	1
sodium fumarate	1

Table A.2: Full organism media compound counts. Listed below is every species in MediaDB and the number of compounds that appear in at least one media formulation for that species (i.e. the union of its media compounds).

Organism	Compound Number
Escherichia coli	81
Lactococcus lactis	65
Leishmania major	63
Shewanella oneidensis	53
Bacillus subtilis	51
Geobacter metallireducens	51
Streptococcus thermophilus	51
Deinococcus radiodurans	43
Lactobacillus plantarum	41
Albidiferax ferrireducens	40
Acinetobacter baylyi	37
Thermotoga maritima	37
Bacillus megaterium	36
Candida glabrata	35
Haemophilus influenzae	35
Streptomyces Coelicolor	33
Ketogulonicigenium vulgare	31
Chromohalobacter salexigens	30
Geobacter sulfurreducens	30
Halobacterium salinarum	30
Saccharomyces cerevisiae	29
Salmonella enterica	29
Yersinia pestis	28
Klebsiella pneumoniae	27
Methanococcus maripaludis	27
Bacillus amyloliquefaciens	25
Mycobacterium tuberculosis	25
Synechocystis PCC6803	25
Aspergillus nidulans	24
Aspergillus terreus	24
Bradyrhizobium japonicum	24
Cyanothece	24
Mannheimia succiniciproducens	24
Pseudomonas putida	23
Francisella tularensis	22
Staphylococcus aureus	22
Clostridium acetobutylicum	20

Table A.2 (cont.)

Organism	Compound Number
Chlamydomonas reinhardtii	18
Hydrogenobacter thermophilus	16
Komagataella pastoris	16
Neiserria meningitidis	16
Pseudomonas aeruginosa	16
Aspergillus niger	15
Clostridium thermocellum	15
Corynebacterium glutamicum	14
Cupriavidus necator	14
Rhizobium etli	14
Natronomonas pharaonis	13
Clostridium beijerinckii	12
Methanosarcina acetivorans	12
Methanosarcina barkeri	12
Neurospora crassa	12
Porphyromonas gingivalis	12
Vibrio vulnificus	11
Aspergillus oryzae	10
Acinetobacter baumannii	9
Zymomonas mobilis	7

Supplementary File A.1: Model simulation on known media. An example Matlab script (growEcoliOnMedia.m) that demonstrates how to simulate growth of a model on media from MediaDB. This script simulates growth of the iJR904 model of *E. coli* on 11 different carbon sources corresponding to 11 different media in MediaDB.

```
function growth_rates = growEcoliOnMedia()

%%%%%%%%%%%%%%%%%%
%There are 11 different media formulations for E.coli
%The default medium contains glucose, ammonium, phosphate, and sulfate
%Each medium contains all of these components, plus a carbon source
%Simulate growth on each medium by turning off the glucose uptake and turning on the proper
carbon source
%
%
%
%Outputs
%growth_rates - an array of the growth rates predicted for maximum biomass yield on each
carbon substrate
%
%Matthew A. Richards 06/10/2014

%Load the iJR904 model
load('iJR904.mat');

%Create array to store growth rates
growth_rates =zeros(11,1);
%Create a cell array of the 11 Carbon Sources
substrates = {'Glucose';'Fructose';'Gluconate';'Acetate';'Pyruvate';...
    'Glycerol';'Succinate';'Lactose';'Malate';'Mannose';'Lactate'};

%Glucose is the default substrate; simulate the model as is
glc_model = model;
solution=optimizeCbModel(glc_model);
growth_rates(1)=solution.f;

%Fructose
frc_model = changeRxnBounds(model,'EX_glc(e)',0,'l');
frc_model = changeRxnBounds(frc_model,'EX_fru(e)',-10,'l');
solution=optimizeCbModel(frc_model);
growth_rates(2)=solution.f;

%Gluconate
glu_model = changeRxnBounds(model,'EX_glc(e)',0,'l');
glu_model = changeRxnBounds(glu_model,'EX_glcn(e)',-10,'l');
solution=optimizeCbModel(glu_model);
growth_rates(3)=solution.f;

%Acetate
ac_model = changeRxnBounds(model,'EX_glc(e)',0,'l');
```

Supplementary File A.1 (cont.)

```
ac_model = changeRxnBounds(ac_model,'EX_ac(e)',-10,'l');
solution=optimizeCbModel(ac_model);
growth_rates(4)=solution.f;

%Pyruvate
pyr_model = changeRxnBounds(model,'EX_glc(e)',0,'l');
pyr_model = changeRxnBounds(pyr_model,'EX_pyr(e)',-10,'l');
solution=optimizeCbModel(pyr_model);
growth_rates(5)=solution.f;

%Glycerol
gly_model = changeRxnBounds(model,'EX_glc(e)',0,'l');
gly_model = changeRxnBounds(gly_model,'EX_glyc(e)',-10,'l');
solution=optimizeCbModel(gly_model);
growth_rates(6)=solution.f;

%Succinate
suc_model = changeRxnBounds(model,'EX_glc(e)',0,'l');
suc_model = changeRxnBounds(suc_model,'EX_succ(e)',-10,'l');
solution=optimizeCbModel(suc_model);
growth_rates(7)=solution.f;

%Lactose
lco_model = changeRxnBounds(model,'EX_glc(e)',0,'l');
lco_model = changeRxnBounds(lco_model,'EX_lcts(e)',-10,'l');
solution=optimizeCbModel(lco_model);
growth_rates(8)=solution.f;

%Malate
mal_model = changeRxnBounds(model,'EX_glc(e)',0,'l');
mal_model = changeRxnBounds(mal_model,'EX_mal-L(e)',-10,'l');
solution=optimizeCbModel(mal_model);
growth_rates(9)=solution.f;

%Mannose
man_model = changeRxnBounds(model,'EX_glc(e)',0,'l');
man_model = changeRxnBounds(man_model,'EX_man(e)',-10,'l');
solution=optimizeCbModel(man_model);
growth_rates(10)=solution.f;

%Lactate
lca_model = changeRxnBounds(model,'EX_glc(e)',0,'l');
lca_model = changeRxnBounds(lca_model,'EX_lac-D(e)',-10,'l');
solution=optimizeCbModel(lca_model);
growth_rates(11)=solution.f;
```

Supplementary File A.1 (cont.)

```
%Print out the responses
fprintf('\n\n');
for i=1:length(growth_rates)
   fprintf('%s: %f\n',substrates{i},growth_rates(i))
end
```

Appendix B: Supporting Information for Chapter 3

Table B.1: iMR540 Reaction Information. A list of every reaction in the iMR540 reconstruction, including the subsystem, ProbAnno likelihood score (if applicable), and the origin tag. The origin tag shows how a particular reaction was added to the reconstruction; "Kbase" reactions were part of the automated Kbase reconstruction and are associated with genes; "Exchange" reactions are non-enzymatic reactions that allow for uptake or secretion of a particular compound by the model; "Physiological" reactions are transport reactions that lack genes but are known to function in the organism; "GapFill" reactions lack any genetic information and were added only to enable growth of the model; "Manual Addition" reactions were added during manual curation and tie to a particular literature reference.

Reaction ID	Subsystem	Likelihood	Origin Tag
rxn02483[c0]	Protocatechuate branch of beta-ketoadipate pathway	0.88062	KBase
rxn00802[c0]	Arginine Biosynthesis	0.9853	KBase
rxn03638[c0]	Peptidoglycan Biosynthesis	0.48895	KBase
rxn06078[c0]	Methionine Biosynthesis	0.50441	KBase
rxn12636[c0]	Protein Degradation	0.96997	KBase
rxn12644[c0]	Protein Degradation	0.96997	KBase
rxn03492[c0]	Coenzyme B12 Biosynthesis	0.9489	KBase
rxn05616[c0]	Transport	0.97285	KBase
rxn02212[c0]	Chorismate Synthesis	0.97879	KBase
rxn03419[c0]	Folate Biosynthesis	0.97092	KBase
rxn03052[c0]	Methionine Biosynthesis	0.94502	KBase
rxn00248[c0]	Glyoxylate Bypass	0.8752	KBase
rxn01302[c0]	Methionine Biosynthesis	0.98208	KBase
rxn00048[c0]	Riboflavin, FMN, and FAD Metabolism	0.94926	KBase
rxn05229[c0]	De Novo Purine Biosynthesis	0.00333	KBase
rxn12637[c0]	Protein Degradation	0.96997	KBase
rxn00321[c0]	Lysine Degradation	0.99311	KBase
rxn07586[c0]	Coenzyme B12 Biosynthesis	0.94767	KBase
rxn00902[c0]	Branched Chain Amino Acid Biosynthesis	0.90003	KBase
rxn00952[c0]	Methionine Biosynthesis	0.50441	KBase
rxn10230[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.98565	KBase
rxn02056[c0]	Heme and Siroheme Biosynthesis	0.80225	KBase
rxn00459[c0]	Glycolysis and Gluconeogenesis	0.99813	KBase
rxn10225[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.15173	KBase
rxn08311[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.15173	KBase
rxn01101[c0]	Serine Biosynthesis	0.98582	KBase
rxn00302[c0]	Folate Biosynthesis	0.97092	KBase

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
rxn05145[c0]	Transport	0.97637	KBase
rxn02988[c0]	NAD/NADP Cofactor Biosynthesis	0.9893	KBase
rxn07307[c0]	Coenzyme M Biosynthesis	0.67663	KBase
rxn02937[c0]	De Novo Purine Biosynthesis	0.98495	KBase
rxn00558[c0]	Glycolysis and Gluconeogenesis	0.9821	KBase
rxn01406[c0]	Polyamine Metabolism	0.98966	KBase
rxn00187[c0]	Glutamine/Glutamate/Aspartate/Asparagine Biosynthesis	0.99374	KBase
rxn01200[c0]	Pentose Phosphate Pathway	0.99527	KBase
rxn05232[c0]	Ribonucleotide Reduction	0.98946	KBase
rxn00835[c0]	Purine Conversions	0.20243	KBase
rxn12844[c0]	Protein Degradation	0.96997	KBase
rxn01678[c0]	Purine Conversions	0.98101	KBase
rxn09208[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.98565	KBase
rxn10060[c0]	NAD/NADP Cofactor Biosynthesis	0.98373	KBase
rxn00100[c0]	Coenzyme A Biosynthesis	0.90055	KBase
rxn01018[c0]	De Novo Pyrimidine Synthesis	0.99141	KBase
rxn00806[c0]	Branched Chain Amino Acid Biosynthesis	0.99416	KBase
rxn02834[c0]	Histidine Biosynthesis	0.95873	KBase
rxn02341[c0]	Coenzyme A Biosynthesis	0.97835	KBase
rxn03175[c0]	Histidine Biosynthesis	0.99177	KBase
rxn00786[c0]	Glycolysis and Gluconeogenesis	0.87243	KBase
rxn01637[c0]	Arginine Biosynthesis	0.94924	KBase
rxn12640[c0]	Protein Degradation	0.96997	KBase
rxn05596[c0]	Transport	0.95817	KBase
rxn02569[c0]	Selenocysteine Metabolism	0.96793	KBase
rxn00710[c0]	De Novo Pyrimidine Synthesis	0.93781	KBase
rxn02775[c0]	Coenzyme B12 Biosynthesis	0.96483	KBase
rxn07589[c0]	Coenzyme B12 Biosynthesis	0.88424	KBase
rxn00717[c0]	Creatine and Creatinine Degradation	0.77609	KBase
rxn10308[c0]	Teichoic and Lipoteichoic Acids Biosynthesis	0.39392	KBase
rxn05249[c0]	Unsaturated Fatty Acid Biosynthesis; Biotin Biosynthesis	0.51238	KBase
rxn02897[c0]	Coenzyme B12 Biosynthesis	0.95937	KBase
rxn03409[c0]	Peptidoglycan Biosynthesis	0.99374	KBase
rxn01301[c0]	Methionine Biosynthesis	0.98208	KBase
rxn07191[c0]	Glycolysis and Gluconeogenesis	0.96817	KBase
rxn04046[c0]	Coenzyme B12 Biosynthesis	0.96483	KBase

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
rxn13782[c0]	Ribosome LSU (Bacteria)	0.00038752	Exchange
rxn03407[c0]	Peptidoglycan Biosynthesis	0.99374	KBase
rxn01682[c0]	Tryptophan Synthesis	0.98441	KBase
rxn02474[c0]	Riboflavin, FMN, and FAD Metabolism	0.94073	KBase
rxn10311[c0]	Teichoic and Lipoteichoic Acids Biosynthesis	0.39392	KBase
rxn01486[c0]	Archaeal Lipids	0.45251	KBase
rxn00898[c0]	Branched Chain Amino Acid Biosynthesis	0.99544	KBase
rxn01069[c0]	Threonine and Homoserine Biosynthesis	0.98565	KBase
rxn02113[c0]	Acetoin, Butanediol Metabolism	0.96586	KBase
rxn00117[c0]	Purine Conversions; Pyrimidine Conversions	0.98101	KBase
rxn00555[c0]	Formaldehyde Assimilation: Ribulose Monophosphate Pathway	0.99786	KBase
rxn00260[c0]	Amino Acid Biosynthesis	0.70583	KBase
rxn02484[c0]	Thiamin Biosynthesis	0.93366	KBase
rxn00527[c0]	Amino Acid Biosynthesis	0.73289	KBase
rxn05171[c0]	Transport	0.85897	KBase
rxn04052[c0]	Coenzyme B12 Biosynthesis	0.9769	KBase
rxn13783[c0]	DNA-Replication	0.11123	Exchange
rxn04045[c0]	Coenzyme B12 Biosynthesis	0.74202	KBase
rxn01974[c0]	Lysine Biosynthesis DAP Pathway	0.97692	KBase
rxn12645[c0]	Protein Degradation	0.96997	KBase
rxn00299[c0]	Folate Biosynthesis	0.97092	KBase
rxn00839[c0]	Purine Conversions; Pyrimidine Conversions	0.98101	KBase
rxn03020[c0]	Methanogenesis	0.94658	KBase
rxn00283[c0]	Alanine Biosynthesis	0.99196	KBase
rxn01022[c0]	Methionine Salvage	0.84713	KBase
rxn10221[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.15173	KBase
rxn09177[c0]	Coenzyme A Biosynthesis	0.9768	KBase
rxn03437[c0]	Branched Chain Amino Acid Biosynthesis	0.99544	KBase
rxn02320[c0]	Histidine Biosynthesis	0.73289	KBase
rxn07587[c0]	Coenzyme B12 Biosynthesis	0.98662	KBase
rxn02507[c0]	Tryptophan Synthesis	0.97169	KBase
rxn00669[c0]	Protein Acetylation and Deacetylation in Bacteria	0.048616	KBase
rxn01333[c0]	Folate Biosynthesis; Pentose Phosphate Pathway	0.99215	KBase
rxn00714[c0]	Pyrimidine Conversions	0.94633	KBase
rxn01434[c0]	Arginine Biosynthesis	0.99306	KBase
rxn01116[c0]	Pentose Phosphate Pathway	0.98379	KBase

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
rxn08308[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.15173	KBase
rxn00409[c0]	Purine Conversions; Pyrimidine Conversions	0.98101	KBase
rxn02213[c0]	Chorismate Synthesis	0.92886	KBase
rxn03062[c0]	Branched Chain Amino Acid Biosynthesis	0.51651	KBase
rxn03446[c0]	Threonine and Homoserine Biosynthesis	0.98565	KBase
rxn01519[c0]	Folate Biosynthesis	0.85571	KBase
rxn00029[c0]	Heme and Siroheme Biosynthesis	0.99335	KBase
rxn03491[c0]	Coenzyme B12 Biosynthesis	0.88424	KBase
rxn01964[c0]	Tryptophan Synthesis	0.98441	KBase
rxn12510[c0]	Coenzyme A Biosynthesis	0.91028	KBase
rxn09207[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.98565	KBase
rxn02380[c0]	Glycolysis and Gluconeogenesis	0.9821	KBase
rxn00474[c0]	Tryptophan Synthesis	0.98441	KBase
rxn03435[c0]	Coenzyme A Biosynthesis; Branched Chain Amino Acid Biosynthesis	0.99321	KBase
rxn01673[c0]	Purine Conversions; Pyrimidine Conversions	0.98101	KBase
rxn12845[c0]	Protein Degradation	0.96997	KBase
rxn00085[c0]	Glutamine/Glutamate/Aspartate/Asparagine Biosynthesis	0.018434	KBase
rxn00832[c0]	De Novo Purine Biosynthesis	0.90124	KBase
rxn00295[c0]	N-Linked Glycosylation in Bacteria	0.90807	KBase
rxn02187[c0]	Valine, Leucine, and Isoleucine Biosynthesis	0.99321	KBase
rxn01219[c0]	Pyrimidine Conversions	0.013479	KBase
rxn03514[c0]	Coenzyme B12 Biosynthesis	0.99463	KBase
rxn01270[c0]	Phenylalanine and Tyrosine Branches from Chorismate	0.61235	KBase
rxn08309[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.15173	KBase
rxn08312[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.15173	KBase
rxn05144[c0]	Pyridoxin(Vitamin B6) Biosynthesis	0.99374	KBase
rxn00410[c0]	Pyrimidine Conversions	0.98711	KBase
rxn05221[c0]	Transport	0.93923	KBase
rxn09210[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.98565	KBase
rxn00782[c0]	Glycolysis and Gluconeogenesis	0.97885	KBase
rxn01607[c0]	Isoprenoid Biosynthesis; Archaeal Lipids	0.88752	KBase
rxn01000[c0]	Phenylalanine and Tyrosine Branches from Chorismate	0.96216	KBase
rxn00675[c0]	Protein Acetylation and Deacetylation in Bacteria	0.95498	KBase

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
rxn02305[c0]	Thiamin Biosynthesis	0.93774	KBase
rxn02296[c0]	Biotin Biosynthesis	0.84669	KBase
rxn12641[c0]	Protein Degradation	0.96997	KBase
rxn10315[c0]	Teichoic and Lipoteichoic Acids Biosynthesis	0.39392	KBase
rxn00711[c0]	De Novo Pyrimidine Synthesis	0.57872	KBase
rxn10309[c0]	Teichoic and Lipoteichoic Acids Biosynthesis	0.39392	KBase
rxn12633[c0]	Protein Degradation	0.96997	KBase
rxn05250[c0]	Unsaturated Fatty Acid Biosynthesis; Biotin Biosynthesis	0.51238	KBase
rxn03174[c0]	Folate Biosynthesis	0.97092	KBase
rxn00763[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.70296	KBase
rxn00138[c0]	NAD/NADP Cofactor Biosynthesis	0.95103	KBase
rxn03136[c0]	De Novo Purine Biosynthesis	0.98898	KBase
rxn05165[c0]	Transport	0.42544	KBase
rxn00988[c0]	Branched Chain Amino Acid Degradation Regulons	0.38842	KBase
rxn05236[c0]	Ribonucleotide Reduction	0.98946	KBase
rxn02835[c0]	Histidine Biosynthesis	0.95605	KBase
rxn00726[c0]	Tryptophan Synthesis	0.78102	KBase
rxn03147[c0]	De Novo Purine Biosynthesis	0.98117	KBase
rxn00650[c0]	Protein Degradation	0.96997	KBase
rxn01241[c0]	TCA Cycle	0.79008	KBase
rxn10229[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.98565	KBase
rxn00776[c0]	tRNA Modification (Bacteria); rRNA Modification (Bacteria)	0.69547	KBase
rxn12646[c0]	Protein Degradation	0.96997	KBase
rxn09296[c0]	Glycine Reductase, Sarcosine Reductase, and Betaine Reductase	0.88497	KBase
rxn01255[c0]	Chorismate Synthesis	0.98883	KBase
rxn03194[c0]	Branched Chain Amino Acid Biosynthesis; Acetoin, Butanediol Metabolism	0.99106	KBase
rxn10222[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.15173	KBase
rxn11938[c0]	Methanogenesis	0.72926	KBase
rxn02287[c0]	Coenzyme B12 Biosynthesis	N/A	KBase
rxn02475[c0]	Riboflavin, FMN, and FAD Metabolism	0.58589	KBase
rxn00147[c0]	Glycolysis and Gluconeogenesis	0.99536	KBase
rxn02774[c0]	Heme and Siroheme Biosynthesis	N/A	KBase
rxn05172[c0]	Transport	0.85897	KBase
rxn03057[c0]	Methionine Salvage	0.97695	KBase

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
rxn01485[c0]	Sialic Acid Metabolism	0.92655	KBase
rxn13784[c0]	tRNA Modification (Bacteria)	0.0056246	Exchange
rxn00245[c0]	None	0.65051	KBase
rxn01629[c0]	Heme and Siroheme Biosynthesis	0.98625	KBase
rxn06874[c0]	Nitrogen Fixation	0.79028	KBase
rxn06299[c0]	Methanogenesis	0.5754	KBase
rxn05313[c0]	Transport	0.93518	KBase
rxn00966[c0]	Ubiquionine Biosynthesis	0	KBase
rxn00785[c0]	Pentose Phosphate Pathway	0.99527	KBase
rxn03084[c0]	De Novo Purine Biosynthesis	0.98232	KBase
rxn00011[c0]	Methionine Degradation	0.99106	KBase
rxn00292[c0]	Sialic Acid Metabolism	0.92219	KBase
rxn05252[c0]	Unsaturated Fatty Acid Biosynthesis; Biotin Biosynthesis	0.51238	KBase
rxn05155[c0]	Transport	0.0081148	KBase
rxn01513[c0]	Pyrimidine Conversions	0.93655	KBase
rxn02476[c0]	Chorismate Synthesis	0.98355	KBase
rxn01575[c0]	Amino Acid Biosynthesis	0.99416	KBase
rxn09206[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.98565	KBase
rxn00781[c0]	Glycolysis and Gluconeogenesis	0.97885	KBase
rxn05181[c0]	Transport	0.54463	KBase
rxn00775[c0]	NAD/NADP Cofactor Biosynthesis	0.73175	KBase
rxn00364[c0]	Pyrimidine Conversions	0.013479	KBase
rxn00515[c0]	Purine Conversions; Pyrimidine Conversions	0.98101	KBase
rxn01025[c0]	Pyrimidine Conversions	0.77609	KBase
rxn02186[c0]	Coenzyme A Biosynthesis; Branched Chain Amino Acid Biosynthesis	0.99321	KBase
rxn10227[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.98565	KBase
rxn12846[c0]	Protein Degradation	0.96997	KBase
rxn05201[c0]	Transport	0.93382	KBase
rxn02789[c0]	Branched Chain Amino Acid Biosynthesis	0.93134	KBase
rxn00830[c0]	Isoprenoid Biosynthesis; Archaeal Lipids	0.98633	KBase
rxn08180[c0]	Biotin Biosynthesis	0.97418	KBase
rxn00800[c0]	De Novo Purine Biosynthesis	0.98898	KBase
rxn05234[c0]	Ribonucleotide Reduction	0.98946	KBase
rxn01465[c0]	De Novo Pyrimidine Synthesis	0.97705	KBase
rxn00863[c0]	Histidine Biosynthesis	0.99292	KBase

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
rxn09449[c0]	Unsaturated Fatty Acid Biosynthesis; Biotin Biosynthesis	0.51238	KBase
rxn03436[c0]	Coenzyme A Biosynthesis; Branched Chain Amino Acid Biosynthesis	0.99321	KBase
rxn08307[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.15173	KBase
rxn05555[c0]	Transport	0.9818	KBase
rxn02159[c0]	Histidine Biosynthesis	0.99292	KBase
rxn00727[c0]	Tryptophan Synthesis	0.78102	KBase
rxn04385[c0]	Coenzyme B12 Biosynthesis	0.56063	KBase
rxn01300[c0]	Methionine Biosynthesis	0.96931	KBase
rxn00175[c0]	Pyruvate Metabolism II: Acetyl-CoA, Acetogenesis from Pyruvate	0.95498	KBase
rxn06979[c0]	Coenzyme B12 Biosynthesis	0.9818	KBase
rxn12642[c0]	Protein Degradation	0.96997	KBase
rxn05247[c0]	Unsaturated Fatty Acid Biosynthesis; Biotin Biosynthesis	0.51238	KBase
rxn00986[c0]	Protein Acetylation and Deacetylation in Bacteria	0.95498	KBase
rxn07588[c0]	Coenzyme B12 Biosynthesis	0.75984	KBase
rxn10473[c0]	Transport	0	Physiological
rxn05559[c0]	Transport	0.53231	KBase
rxn03068[c0]	Coenzyme A Biosynthesis; Branched Chain Amino Acid Biosynthesis	0.99321	KBase
rxn12634[c0]	Protein Degradation	0.96997	KBase
rxn01269[c0]	Chorismate Synthesis	0.69666	KBase
rxn04048[c0]	Coenzyme B12 Biosynthesis	0.99463	KBase
rxn00278[c0]	Pyruvate Alanine Serine Interconversions	0.9872	KBase
rxn00790[c0]	De Novo Purine Biosynthesis	0.99413	KBase
rxn00060[c0]	Heme and Siroheme Biosynthesis	0.972	KBase
rxn04050[c0]	Coenzyme B12 Biosynthesis	0.9489	KBase
rxn02473[c0]	Histidine Biosynthesis	0.99173	KBase
rxn00834[c0]	Purine Conversions	0.96349	KBase
rxn12639[c0]	Protein Degradation	0.96997	KBase
rxn03080[c0]	Riboflavin, FMN, and FAD Metabolism	0.96251	KBase
rxn03085[c0]	Methanogenesis	0.94026	KBase
rxn02297[c0]	Sphingolipid Biosynthesis; Biotin Biosynthesis	0.59355	KBase
rxn05040[c0]	Riboflavin, FMN, and FAD Metabolism	0.67922	KBase
rxn00105[c0]	NAD/NADP Cofactor Biosynthesis	0.95417	KBase
rxn00799[c0]	TCA Cycle	0.68542	KBase
rxn00313[c0]	Lysine Biosynthesis DAP Pathway	0.98454	KBase

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
rxn10226[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.98565	KBase
rxn05248[c0]	Unsaturated Fatty Acid Biosynthesis; Biotin Biosynthesis	0.51238	KBase
rxn10223[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.15173	KBase
rxn00285[c0]	TCA Cycle	0.98828	KBase
rxn10307[c0]	Teichoic and Lipoteichoic Acids Biosynthesis	0.39392	KBase
rxn00770[c0]	Pentose Phosphate Pathway; De Novo Purine Biosynthesis	0.97264	KBase
rxn00192[c0]	Arginine Biosynthesis	0.9197	KBase
rxn05938[c0]	Pyruvate:Ferredoxin Oxidoreductase	0.95261	KBase
rxn00211[c0]	Sucrose Metabolism; Teichuronic Acid Biosynthesis	0.84935	KBase
rxn04783[c0]	De Novo Purine Biosynthesis	0.98317	KBase
rxn01739[c0]	Chorismate Synthesis	0.93	KBase
rxn00543[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.70296	KBase
rxn05235[c0]	Ribonucleotide Reduction	0.98946	KBase
rxn06591[c0]	Heme and Siroheme Biosynthesis	0.98094	KBase
rxn00126[c0]	Methionine Biosynthesis	0.98957	KBase
rxn00414[c0]	De Novo Pyrimidine Synthesis	0.99921	KBase
rxn06937[c0]	tRNA Aminoacylation; Heme and Siroheme Biosynthesis	0.60018	KBase
rxn03127[c0]	Methanogenesis	0.9704	KBase
rxn10312[c0]	Teichoic and Lipoteichoic Acids Biosynthesis	0.39392	KBase
rxn02508[c0]	Tryptophan Synthesis	0.92342	KBase
rxn00127[c0]	Polyamine Metabolism	0.95061	KBase
rxn05251[c0]	Unsaturated Fatty Acid Biosynthesis; Biotin Biosynthesis	0.51238	KBase
rxn02938[c0]	De Novo Purine Biosynthesis	0.98415	KBase
rxn00139[c0]	cAMP Signaling in Bacteria; Purine Conversions	0.98654	KBase
rxn05736[c0]	Unsaturated Fatty Acid Biosynthesis; Biotin Biosynthesis	0.51238	KBase
rxn05527[c0]	Transport	0.76394	KBase
rxn02465[c0]	Arginine Biosynthesis	0.9949	KBase
rxn09205[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.98565	KBase
rxn05957[c0]	Methionine Biosynthesis	0	KBase
rxn03513[c0]	Coenzyme B12 Biosynthesis	0.97411	KBase
rxn00549[c0]	Glycolysis and Gluconeogenesis	0.97936	KBase
rxn01257[c0]	Folate Biosynthesis; Tryptophan Synthesis	0.78102	KBase
rxn01740[c0]	Chorismate Synthesis	0.82226	KBase

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
rxn12847[c0]	Protein Degradation	0.96997	KBase
rxn00416[c0]	Glutamine/Glutamate/Aspartate/Asparagine Biosynthesis	0.9234	KBase
rxn01620[c0]	L-Fucose Utilization	0.15504	KBase
rxn00903[c0]	Amino Acid Biosynthesis	0.99416	KBase
rxn00947[c0]	HMG CoA Synthesis	0.51238	KBase
rxn10310[c0]	Teichoic and Lipoteichoic Acids Biosynthesis	0.39392	KBase
rxn00838[c0]	Histidine Biosynthesis; Purine Conversions	0.98617	KBase
rxn00297[c0]	Sialic Acid Metabolism	0.92219	KBase
rxn00065[c0]	cAMP Signaling in Bacteria	0.88581	KBase
rxn00141[c0]	Methionine Biosynthesis/Degradation	0.99086	KBase
rxn01454[c0]	Isoprenoid Biosynthesis; Archaeal Lipids	0.97027	KBase
rxn00337[c0]	Lysine Biosynthesis DAP Pathway; Threonine and Homoserine Biosynthesis	0.71791	KBase
rxn00777[c0]	Pentose Phosphate Pathway	0.98497	KBase
rxn01512[c0]	Purine Conversions; Pyrimidine Conversions	0.98101	KBase
rxn01466[c0]	Archaeal Lipids	0.12187	KBase
rxn02185[c0]	Branched Chain Amino Acid Synthesis; Acetoin, Butanediol Metabolism	0.99106	KBase
rxn03406[c0]	Peptidoglycan Biosynthesis; Amino Acid Biosynthesis	0.99374	KBase
rxn01106[c0]	Glycolysis and Gluconeogenesis; Theronine and Homoserine Biosynthesis	0.6879	KBase
rxn00293[c0]	Peptidoglycan Biosynthesis	0.48895	KBase
rxn05740[c0]	Glycogen Metabolism	0.96109	KBase
rxn11544[c0]	Coenzyme B12 Biosynthesis	0.9818	KBase
rxn02277[c0]	Biotin Biosynthesis	0.95841	KBase
rxn12638[c0]	Protein Degradation	0.96997	KBase
rxn02480[c0]	Methanogenesis	0.97999	KBase
rxn12643[c0]	Protein Degradation	0.96997	KBase
rxn03079[c0]	Methanogenesis	0.96121	KBase
rxn02264[c0]	Heme and Siroheme Biosynthesis	0.92227	KBase
rxn05209[c0]	Transport	0.2944	KBase
rxn08306[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.15173	KBase
rxn00148[c0]	Glycolysis and Gluconeogenesis	0.99132	KBase
rxn00214[c0]	N-Linked Glycosylation in Bacteria	0.90807	KBase
rxn05197[c0]	Transport	0.2031	KBase
rxn05289[c0]	Thioredoxin-disulfide Reductase; Pyrimidine Conversions	0.95363	KBase
rxn10228[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.98565	KBase

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
rxn01977[c0]	Glycolysis and Gluconeogenesis	0.9821	KBase
rxn09448[c0]	Unsaturated Fatty Acid Biosynthesis; Biotin Biosynthesis	0.51238	KBase
rxn03421[c0]	Folate Biosynthesis	0.97092	KBase
rxn12635[c0]	Protein Degradation	0.96997	KBase
rxn00213[c0]	Sucrose Metabolism	0.98642	KBase
rxn10314[c0]	Teichoic and Lipoteichoic Acids Biosynthesis	0.39392	KBase
rxn00493[c0]	Amino Acid Biosynthesis	0.61235	KBase
rxn00789[c0]	Histidine Biosynthesis	0.98209	KBase
rxn06077[c0]	Peptide Methionine Sufoxide Reductase	0.95249	KBase
rxn02895[c0]	De Novo Purine Biosynthesis	0.99089	KBase
rxn09211[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.98565	KBase
rxn02312[c0]	Biotin Biosynthesis	0.94916	KBase
rxn00438[c0]	Thiamin Biosynthesis	0.9588	KBase
rxn10474[c0]	Transport	0.97285	KBase
rxn01643[c0]	Lysine Biosynthesis; Threonine and Homoserine Biosynthesis	0.98215	KBase
rxn09209[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.98565	KBase
rxn02431[c0]	Methanogenesis	0.94866	KBase
rxn01999[c0]	None	0	KBase
rxn01362[c0]	De Novo Pyrimidine Synthesis	0.9532	KBase
rxn06493[c0]	Glycine and Serine Utilization	0.79008	KBase
rxn02402[c0]	NAD/NADP Cofactor Biosynthesis	0.96992	KBase
rxn10224[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.15173	KBase
rxn10231[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.98565	KBase
rxn02811[c0]	Branched Chain Amino Acid Biosynthesis	0.93134	KBase
rxn05939[c0]	TCA Cycle	0	KBase
rxn01019[c0]	Arginine Biosynthesis	0.9882	KBase
rxn00412[c0]	Pyrimidine Conversions	0.98711	KBase
rxn08310[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.15173	KBase
rxn03135[c0]	Histidine Biosynthesis	0.99701	KBase
rxn00469[c0]	Arginine Biosynthesis	0.7241	KBase
rxn12512[c0]	Coenzyme A Biosynthesis	0.9768	KBase
rxn00791[c0]	Tryptophan Synthesis	0.98709	KBase
rxn08040[c0]	Teichoic and Lipoteichoic Acids Biosynthesis	0.39392	KBase
rxn01636[c0]	Arginine Biosynthesis	0.988	KBase

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
rxn00914[c0]	Purine Conversions	0.20243	KBase
rxn03843[c0]	Isoprenoid Biosynthesis	0.40242	KBase
rxn01329[c0]	Mannose Metabolism	0.84904	KBase
rxn04047[c0]	Coenzyme B12 Biosynthesis	0.97411	KBase
rxn05177[c0]	Transport	0.031108	KBase
rxn01353[c0]	Purine Conversions; Pyrimidine Conversions	0.98101	KBase
rxn00077[c0]	NAD/NADP Cofactor Biosynthesis	0.98373	KBase
rxn05667[c0]	Transport	0	Physiological
rxn09450[c0]	Unsaturated Fatty Acid Biosynthesis; Biotin Biosynthesis	0.51238	KBase
rxn01213[c0]	Archaeal Lipids	0.12187	KBase
rxn01917[c0]	Arginine Biosynthesis	0.99531	KBase
rxn01100[c0]	Glycolysis and Gluconeogenesis	0.98697	KBase
rxn02339[c0]	Amino Acid Biosynthesis	0.70583	KBase
rxn02175[c0]	Coenzyme A Biosynthesis	0.87611	KBase
rxn00288[c0]	Succinate Dehydrogenase	0.59016	KBase
rxn01790[c0]	Coenzyme A Biosynthesis; Branched Chain Amino Acid Biosynthesis	0.99321	KBase
rxn03075[c0]	Thiamin Biosynthesis	0.98662	KBase
rxn00747[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.95959	KBase
rxn00707[c0]	Pyrimidine Conversions	0.013479	KBase
rxn03108[c0]	Thiamin Biosynthesis	0.93366	KBase
rxn03512[c0]	Coenzyme B12 Synthesis	0.9769	KBase
rxn00237[c0]	Purine Conversions; Pyrimidine Conversions	0.98101	KBase
rxn00420[c0]	Glycine and Serine Utilization; Serine Biosynthesis	0.9911	KBase
rxn00001[c0]	HPr Catabolite Repression System	0.9822	KBase
rxn10313[c0]	Teichoic and Lipoteichoic Acids Biosynthesis	0.39392	KBase
rxn01268[c0]	Chorismate Synthesis	0.69666	KBase
rxn10220[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.15173	KBase
rxn00858[c0]	Polyamine Metabnolism; Arginine and Ornithine Degradation	0.98397	KBase
rxn10571[c0]	Transport	N/A	Physiological
rxn05195[c0]	Transport	N/A	Physiological
rxn00062[c0]	Non-Growth Associated Maintenance	N/A	Gapfill
rxn01208[c0]	None	N/A	Gapfill
rxn05319[c0]	Transport	N/A	Physiological
rxn05467[c0]	Transport	N/A	Physiological
rxn03393[c0]	Hydroxyaromatic Decarboxylase Family	0.96944	KBase

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
rxn00623[c0]	Sulfate Assimilation	0.004751	KBase
rxn02832[c0]	Menaquinone and Phylloquinone Biosynthesis	0	Gapfill
rxn10052[c0]	Purine Conversions	0	Gapfill
rxn05388[c0]	Fatty Acid Biosynthesis	0.012453	KBase
rxn05359[c0]	Fatty Acid Biosynthesis	0.012453	KBase
rxn02288[c0]	Heme and Siroheme Biosynthesis	0.91453	KBase
rxn08766[c0]	None	0.0027405	KBase
rxn08019[c0]	None	N/A	Gapfill
rxn00737[c0]	Branched Chain Amino Acid Biosynthesis	0	Gapfill
rxn05396[c0]	Fatty Acid Biosynthesis	0.012453	KBase
rxn05029[c0]	Coenzyme B12 Biosynthesis	0	Gapfill
rxn10181[c0]	Transport	0	Physiological
rxn01675[c0]	dTDP-Rhamnose Synthesis	0.19741	KBase
rxn01644[c0]	Lysine Biosynthesis DAP Pathway	0.021161	KBase
rxn01997[c0]	dTDP-Rhamnose Synthesis	0.015835	KBase
rxn05379[c0]	Fatty Acid Biosynthesis	0.012453	KBase
rxn05371[c0]	Fatty Acid Biosynthesis	0.012453	KBase
rxn00300[c0]	Riboflavin, FMN, and FAD Metabolism	0.39555	KBase
rxn00392[c0]	Ubiquionine Biosynthesis	0	Gapfill
rxn03397[c0]	Ubiquionine Biosynthesis	0.68382	KBase
rxn10481[c0]	Transport	0.92767	KBase
rxn13477[c0]	None	N/A	Gapfill
rxn10954[c0]	None	N/A	Gapfill
rxn02914[c0]	Glycine and Serine Utilization; Serine Biosynthesis; Pyridoxin Biosynthesis	0.0098046	KBase
rxn00350[c0]	Glutathione: Biosynthesis and gamma-Glutamyl Cycle	0	Gapfill
rxn03536[c0]	Coenzyme B12 Biosynthesis	0	Gapfill
rxn05054[c0]	Coenzyme B12 Biosynthesis	0.8679	KBase
rxn02160[c0]	Histidine Biosynthesis	0.2019	KBase
rxn12008[c0]	Polyprenyl Diphosphate Biosynthesis	0.12187	KBase
rxn05039[c0]	Riboflavin, FMN, and FAD Metabolism	0	Gapfill
rxn02269[c0]	None	N/A	Gapfill
rxn03891[c0]	Polyprenyl Diphosphate Biosynthesis	0.12187	KBase
rxn01256[c0]	Chorismate Synthesis; Phenylalanine Synthesis	0	Gapfill
rxn00102[c0]	CO2 Uptake, Carboxysome	0.91831	KBase
rxn09562[c0]	Purine Conversions	0	Gapfill
rxn01972[c0]	Lysine Biosynthesis DAP Pathway	0	Gapfill
rxn05513[c0]	Transport	N/A	Physiological

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
rxn03537[c0]	Coenzyme B12 Biosynthesis	0	Gapfill
rxn03030[c0]	Lysine Biosynthesis DAP Pathway	0.8817	KBase
rxn00119[c0]	None	0.81695	KBase
rxn01265[c0]	NAD/NADP Cofactor Biosynthesis	0	Gapfill
rxn03919[c0]	LOS Core Oligosaccharide Biosynthesis	0.033534	KBase
rxn01068[c0]	Threonine Degradation; Glycine Biosynthesis	0	Gapfill
rxn05384[c0]	Fatty Acid Biosynthesis	0.012453	KBase
rxn02831[c0]	Menaquinone and Phylloquinone Biosynthesis	0	Gapfill
rxn11650[c0]	Coenzyme B12 Biosynthesis	0.75905	KBase
rxn00137[c0]	Inorganic Sulfur Assimilation	0	Gapfill
rxn03150[c0]	Coenzyme B12 Biosynthesis	0	Gapfill
rxn00172[c0]	Pyruvate Metabolism II: Acetyl-CoA, Acetogenesis from Pyruvate	0.020916	KBase
rxn00917[c0]	Purine Conversions	0.89297	KBase
rxn00274[c0]	Glycine and Serine Utilization; Threonine Degradation; Glycine Biosynthesis	0.39144	KBase
rxn05400[c0]	Fatty Acid Biosynthesis	0.012453	KBase
rxn03540[c0]	Coenzyme B12 Biosynthesis	0.81527	KBase
rxn01538[c0]	Thiamin Biosynthesis	0	Gapfill
rxn05404[c0]	Fatty Acid Biosynthesis	0.012453	KBase
rxn05392[c0]	Fatty Acid Biosynthesis	0.012453	KBase
rxn05367[c0]	Fatty Acid Biosynthesis	0.012453	KBase
rxn03408[c0]	Peptidoglycan Biosynthesis	0.59921	KBase
rxn06729[c0]	KDO2-Lipid A Biosynthesis	0.0012682	KBase
rxn00470[c0]	Polyamine Metabolism; Arginine and Ornithine Degradation	0	Gapfill
rxn06023[c0]	Fatty Acid Biosynthesis	0	Gapfill
rxn00086[c0]	Glutathione: Redox Cycle	0.034515	KBase
rxn03086[c0]	Lysine Biosynthesis DAP Pathway	0.17646	KBase
rxn01332[c0]	Chorismate Synthesis	0.0087099	KBase
rxn03904[c0]	Peptidoglycan Biosynthesis	0.080091	KBase
rxn08618[c0]	LOS core oligosaccharide biosynthesis	0.0097104	KBase
rxn00833[c0]	None	N/A	Gapfill
rxn01258[c0]	Menaquinone and Phylloquinone Biosynthesis	0	Gapfill
rxn00611[c0]	Isoprenoid Biosynthesis; Archaeal Lipids	0	Gapfill
rxn00122[c0]	Riboflavin, FMN, and FAD Metabolism	0	Gapfill
rxn01117[c0]	KDO2-Lipid A Biosynthesis	0.23112	KBase
rxn04675[c0]	None	N/A	Gapfill
rxn05735[c0]	Glycerolipid and Glycerophospholipid Metabolism (Bacteria)	0.26823	KBase

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
rxn02929[c0]	Lysine Biosynthesis DAP Pathway	0.34429	KBase
rxn05375[c0]	Fatty Acid Biosynthesis	0.012453	KBase
rxn09128[c0]	Triacylglycerol Metabolism	0.0016864	KBase
rxn05363[c0]	Fatty Acid Biosynthesis	0.012453	KBase
rxn02898[c0]	Menaquinone and Phylloquinone Biosynthesis	0	Gapfill
rxn03538[c0]	Coenzyme B12 Biosynthesis; Heme and Siroheme Biosynthesis	0.86044	KBase
rxn08349[c0]	None	N/A	Gapfill
rxn05023[c0]	None	0	Gapfill
rxn00471[c0]	Threonine Degradation; Arginine and Ornithine Degradation	0	Gapfill
rxn09429[c0]	None	N/A	Gapfill
rxn00346[c0]	Coenzyme A Biosynthesis	0	Gapfill
rxn00351[c0]	Glutathione Biosynthesis	0.025131	KBase
rxn01791[c0]	Coenzyme A Biosynthesis	0.0038642	KBase
rxn01501[c0]	Isoprenoid Biosynthesis; Archaeal Lipids	0.96521	KBase
rxn00250[c0]	Pyruvate Metabolism I: Anaplerotic Reactions, PEP	0.34522	KBase
rxn00646[c0]	Glutathione Biosynthesis	0.14345	KBase
rxn05466[c0]	Transport	0.84551	KBase
rxn09433[c0]	None	N/A	Gapfill
rxn05287[c0]	Polyprenyl Diphosphate Biosynthesis	0.13405	KBase
rxn11703[c0]	Menaquinone and Phylloquinone Biosynthesis	0	Gapfill
rxn05108[c0]	Methionine Salvage	0.093935	KBase
rxn11702[c0]	Menaquinone and Phylloquinone Biosynthesis	0	Gapfill
rxn05744[c0]	None	N/A	Gapfill
rxn03892[c0]	Polyprenyl Diphosphate Biosynthesis	0.13405	KBase
rxn02155[c0]	NAD/NADP Cofactor Biosynthesis	0	Gapfill
rxn00258[c0]	None	N/A	Gapfill
rxn05104[c0]	Methionine Salvage	0.035948	KBase
rxn05909[c0]	None	N/A	Gapfill
rxn00157[c0]	Formate Hydrogenase	0.030184	KBase
rxn00178[c0]	Branched Chain Amino Acid Degradation Regulons	0.39658	KBase
rxn05024[c0]	Menaquinone and Phylloquinone Biosynthesis	0	Gapfill
rxn00134[c0]	Purine Conversions	0.32194	KBase
biomass0	Exchange	N/A	Physiological
EX_cpd00254[e0]	Exchange	N/A	Exchange
EX_cpd00009[e0]	Exchange	N/A	Exchange
EX_cpd00067[e0]	Exchange	N/A	Exchange

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
EX_cpd00205[e0]	Exchange	N/A	Exchange
EX_cpd00209[e0]	Exchange	N/A	Exchange
EX_cpd00971[e0]	Exchange	N/A	Exchange
EX_cpd00129[e0]	Exchange	N/A	Exchange
EX_cpd00210[e0]	Exchange	N/A	Exchange
EX_cpd00053[e0]	Exchange	N/A	Exchange
EX_cpd00540[e0]	Exchange	N/A	Exchange
EX_cpd00226[e0]	Exchange	N/A	Exchange
EX_cpd10515[e0]	Exchange	N/A	Exchange
EX_cpd00099[e0]	Exchange	N/A	Exchange
EX_cpd00047[e0]	Exchange	N/A	Exchange
EX_cpd00307[e0]	Exchange	N/A	Exchange
EX_cpd15302[c0]	Exchange	N/A	Exchange
EX_cpd00092[e0]	Exchange	N/A	Exchange
EX_cpd00149[e0]	Exchange	N/A	Exchange
EX_cpd00305[e0]	Exchange	N/A	Exchange
EX_cpd00073[e0]	Exchange	N/A	Exchange
EX_cpd10516[e0]	Exchange	N/A	Exchange
EX_cpd00001[e0]	Exchange	N/A	Exchange
EX_cpd00011[e0]	Exchange	N/A	Exchange
EX_cpd11416[c0]	Exchange	N/A	Exchange
EX_cpd00034[e0]	Exchange	N/A	Exchange
EX_cpd01741[e0]	Exchange	N/A	Exchange
EX_cpd00355[e0]	Exchange	N/A	Exchange
EX_cpd00058[e0]	Exchange	N/A	Exchange
EX_cpd00558[e0]	Exchange	N/A	Exchange
EX_cpd00030[e0]	Exchange	N/A	Exchange
EX_cpd00063[e0]	Exchange	N/A	Exchange
EX_cpd00655[e0]	Exchange	N/A	Exchange
EX_cpd15269[e0]	Exchange	N/A	Exchange
EX_cpd03422[e0]	Exchange	N/A	Exchange
EX_cpd01080[e0]	Exchange	N/A	Exchange
EX_cpd00111[e0]	Exchange	N/A	Exchange
EX_cpd03847[e0]	Exchange	N/A	Exchange
EX_cpd00013[e0]	Exchange	N/A	Exchange
EX_cpd01024[e0]	Exchange	N/A	Manual Addition
EX_cpd11640[e0]	Exchange	N/A	Manual Addition

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
HdrABC	Methanogenesis	N/A	Manual Addition
Eha/Ehb	Methanogenesis	N/A	Manual Addition
rxn04042[c0]	Glycolysis	N/A	Manual Addition
rxn04043[c0]	Glycolysis	N/A	Manual Addition
rxn04026[c0]	Coenzyme M Biosynthesis	N/A	Manual Addition
rxn04934[c0]	Coenzyme M Biosynthesis	N/A	Manual Addition
rxn04036[c0]	Coenzyme M Biosynthesis	N/A	Manual Addition
rxn07741[c0]	None	N/A	Manual Addition
rxn05109[c0]	None	N/A	Manual Addition
rxn02749[c0]	None	N/A	Manual Addition
rxn02751[c0]	None	N/A	Manual Addition
rxn00735[c0]	None	N/A	Manual Addition
rxn08043[c0]	None	N/A	Manual Addition
rxn08764[c0]	None	N/A	Manual Addition
rxn00405[c0]	Arginine and Ornithine Degradation	0.0051898	Manual Addition
EX_cpd00029[e0]	Exchange	N/A	Manual Addition
rxn10904[c0]	Transport	N/A	Manual Addition
rxn10561[c0]	Amino Acid Biosynthesis	N/A	Manual Addition
rxn06696[c0]	Methanogenesis	N/A	Manual Addition
ATPS	Methanogenesis	N/A	Manual Addition
Fdh	Methanogenesis	N/A	Manual Addition
Hdr_formate	Methanogenesis	N/A	Manual Addition
rxn00249[c0]	None	N/A	Manual Addition
Tfr	None	N/A	Manual Addition
EX_cpd00035[e0]	Exchange	N/A	Manual Addition

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
EX_cpd00117[e0]	Exchange	N/A	Manual Addition
rxn05215[c0]	Transport	0.0012794	Manual Addition
rxn13660[c0]	Transport	N/A	Manual Addition
EX_cpd00528[e0]	Exchange	N/A	Manual Addition
rxn10577[c0]	Transport	N/A	Manual Addition
rxn10541[c0]	Transport	N/A	Manual Addition
EX_cpd00239[e0]	Exchange	N/A	Manual Addition
Dsr-LP	Sulfur Assimilation	N/A	Manual Addition
rxn10434[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition
rxn10608[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition
rxn10599[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition
rxn10472[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition
rxn10612[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition
rxn10433[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition
rxn10595[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition
rxn10468[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition
rxn10610[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition
rxn10435[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition
rxn10596[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition
rxn10469[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition
rxn10611[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition
rxn11855[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition
rxn10424[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition
rxn10425[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition
rxn10475[c0]	Coenzyme B Biosynthesis	N/A	Manual Addition

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
MptA	Tetrahydromethanopterin Biosynthesis	N/A	Manual Addition
rxn10490[c0]	Tetrahydromethanopterin Biosynthesis	N/A	Manual Addition
rxn03168[c0]	Tetrahydromethanopterin Biosynthesis	0	Manual Addition
rxn02504[c0]	Tetrahydromethanopterin Biosynthesis	0.4811	Manual Addition
rxn02503[c0]	Tetrahydromethanopterin Biosynthesis	0.00024841	Manual Addition
rxn10446[c0]	Tetrahydromethanopterin Biosynthesis	N/A	Manual Addition
rxn10491[c0]	Tetrahydromethanopterin Biosynthesis	N/A	Manual Addition
H4MPTs	Tetrahydromethanopterin Biosynthesis	N/A	Manual Addition
rxn10432[c0]	Tetrahydromethanopterin Biosynthesis	N/A	Manual Addition
ADTHs	Tetrahydromethanopterin Biosynthesis	N/A	Manual Addition
ADTHOR	Tetrahydromethanopterin Biosynthesis	N/A	Manual Addition
3DHQAT	Tetrahydromethanopterin Biosynthesis	N/A	Manual Addition
4ADSs	Tetrahydromethanopterin Biosynthesis	N/A	Manual Addition
4ASDH	Tetrahydromethanopterin Biosynthesis	N/A	Manual Addition
4ASDHT	Tetrahydromethanopterin Biosynthesis	N/A	Manual Addition
ABEEs	Tetrahydromethanopterin Biosynthesis	N/A	Manual Addition
rxn00979[c0]	Glycolate, Glyoxylate Interconversions	0.064954	Manual Addition
rxn00512[c0]	Glycolate, Glyoxylate Interconversions	0.06606	Manual Addition
rxn00272[c0]	Serine-Glyoxylate Cycle	0.25782	Manual Addition
COMs	Coenzyme M Biosynthesis	N/A	Manual Addition
rxn00529[c0]	Methanofuran Biosynthesis	0.11606	Manual Addition
MfnD	Methanofuran Biosynthesis	N/A	Manual Addition
MfnB	Methanofuran Biosynthesis	N/A	Manual Addition
MfnC	Methanofuran Biosynthesis	N/A	Manual Addition
F1Pp	Methanofuran Biosynthesis	N/A	Manual Addition

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
F1PPc	Methanofuran Biosynthesis	N/A	Manual Addition
MFs	Methanofuran Biosynthesis	N/A	Manual Addition
rxn10508[c0]	Coenzyme F430 Biosynthesis	N/A	Manual Addition
rxn10509[c0]	Coenzyme F430 Biosynthesis	N/A	Manual Addition
rxn10510[c0]	Coenzyme F430 Biosynthesis	N/A	Manual Addition
rxn10511[c0]	Coenzyme F430 Biosynthesis	N/A	Manual Addition
rxn10512[c0]	Coenzyme F430 Biosynthesis	N/A	Manual Addition
EX_cpd00244[e0]	Exchange	N/A	Manual Addition
rxn05174[c0]	Transport	0	Manual Addition
rxn00097[c0]	Purine Conversions	0	Manual Addition
rxn10499[c0]	Coenzyme F420 Biosynthesis	N/A	Manual Addition
rxn01053[c0]	Coenzyme F420 Biosynthesis	0.10008	Manual Addition
rxn10567[c0]	Coenzyme F420 Biosynthesis	N/A	Manual Addition
rxn10420[c0]	Coenzyme F420 Biosynthesis	N/A	Manual Addition
rxn10566[c0]	Coenzyme F420 Biosynthesis	N/A	Manual Addition
rxn10525[c0]	Coenzyme F420 Biosynthesis	N/A	Manual Addition
rxn10526[c0]	Coenzyme F420 Biosynthesis	N/A	Manual Addition
rxn10527[c0]	Coenzyme F420 Biosynthesis	N/A	Manual Addition
rxn05734[c0]	Methylglyoxal Metabolism	0.26823	Manual Addition
rxn01361[c0]	De Novo Pyrimidine Synthesis	N/A	Manual Addition
CODH	None	N/A	Manual Addition
ACS	None	N/A	Manual Addition
rxn10480[c0]	Transport	N/A	Manual Addition
EX_cpd00204[e0]	Exchange	N/A	Manual Addition
FNO	None	N/A	Manual Addition

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
EX_cpd00131[e0]	Exchange	N/A	Manual Addition
Mot	Transport	N/A	Manual Addition
HcyS	Methionine Biosynthesis	N/A	Manual Addition
MetS	Methionine Biosynthesis	N/A	Manual Addition
rxn02430[c0]	None	N/A	Manual Addition
H4MPT3M2Om	Coenzyme A Biosynthesis	N/A	Manual Addition
H4MPTdUMPm	Pyrimidine Conversions	N/A	Manual Addition
FH4MPTAf	De Novo Purine Biosynthesis	N/A	Manual Addition
HPAr	Coenzyme F420 Biosynthesis	N/A	Manual Addition
rxn02377[c0]	None	0	Manual Addition
rxn00298[c0]	Archaellin Synthesis	N/A	Manual Addition
UGAor	Archaellin Synthesis	N/A	Manual Addition
UGNAa	Archaellin Synthesis	N/A	Manual Addition
UGNAna	Archaellin Synthesis	N/A	Manual Addition
UGNAe	Archaellin Synthesis	N/A	Manual Addition
UMNAat	Archaellin Synthesis	N/A	Manual Addition
2NACmt	Archaellin Synthesis	N/A	Manual Addition
TSot	Archaellin Synthesis	N/A	Manual Addition
GLCgt	Archaellin Synthesis	N/A	Manual Addition
MANgt	Archaellin Synthesis	N/A	Manual Addition
2NACgt	Archaellin Synthesis	N/A	Manual Addition
TSost	Archaellin Synthesis	N/A	Manual Addition
TSf	Transport	N/A	Manual Addition
GALgt	Archaellin Synthesis	N/A	Manual Addition
EX_NAC[c0]	Exchange	N/A	Manual Addition

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
EX_Membrane_lipid[c0]	Exchange	N/A	Manual Addition
EX_Flagellin[e0]	Exchange	N/A	Manual Addition
rxn10560[c0]	Transport	N/A	Manual Addition
rxn10559[c0]	Transport	N/A	Manual Addition
rxn10587[c0]	Transport	N/A	Manual Addition
rxn10586[c0]	Amino Acid Biosynthesis	N/A	Manual Addition
rxn01946[c0]	Amino Acid Biosynthesis	N/A	Manual Addition
rxn01842[c0]	Amino Acid Biosynthesis	0.043068	Manual Addition
rxn10563[c0]	Amino Acid Biosynthesis	N/A	Manual Addition
rxn10562[c0]	Amino Acid Biosynthesis	N/A	Manual Addition
rxn00483[c0]	Amino Acid Biosynthesis	0	Manual Addition
EX_cpd00430[e0]	Exchange	N/A	Manual Addition
EX_cpd00489[e0]	Exchange	N/A	Manual Addition
EX_cpd00703[e0]	Exchange	N/A	Manual Addition
rxn05166[c0]	Transport	N/A	Manual Addition
rxn10447[c0]	Transport	N/A	Manual Addition
rxn01491[c0]	Methylglyoxal Metabolism	N/A	Manual Addition
F6PG3Pl	Methylglyoxal Metabolism	N/A	Manual Addition
2OPs	Methylglyoxal Metabolism	N/A	Manual Addition
DKFPs1	Methylglyoxal Metabolism	N/A	Manual Addition
DKFPs2	Methylglyoxal Metabolism	N/A	Manual Addition
rxn01618[c0]	Methylglyoxal Metabolism	0	Manual Addition
LAFor	Methylglyoxal Metabolism	N/A	Manual Addition
GLYPs	Methylglyoxal Metabolism	N/A	Manual Addition
rxn00149[c0]	None	N/A	Manual Addition

Table B.1 (cont.)

Reaction ID	Subsystem	Likelihood	Origin Tag
rxn00150[c0]	None	N/A	Manual Addition
rxn09249[c0]	Selenocysteine metabolism	N/A	Manual Addition
rxn11751[c0]	Selenocysteine metabolism	N/A	Manual Addition
rxn11752[c0]	Selenocysteine metabolism	N/A	Manual Addition
rxn11950[c0]	Selenocysteine metabolism	N/A	Manual Addition
rxn11571[c0]	Selenocysteine metabolism	N/A	Manual Addition
rxn07210[c0]	Selenocysteine metabolism	N/A	Manual Addition
SELt	Transport	N/A	Manual Addition
EX_cpd03396[e0]	Exchange	N/A	Manual Addition
EX_cpd15573[c0]	Exchange	N/A	Manual Addition
rxn01217[c0]	None	0	Manual Addition
rxn00915[c0]	Purine conversions	0.0022985	Manual Addition
rxn00836[c0]	Purine conversions	0.0022985	Manual Addition
rxn13772[c0]	Isoprenoid Biosynthesis; Archaeal Lipids	N/A	Manual Addition
IPk	Isoprenoid Biosynthesis; Archaeal Lipids	N/A	Manual Addition
rxn11998[c0]	Archaeal Lipids	N/A	Manual Addition
rxn03114[c0]	Archaeal Lipids	N/A	Manual Addition
rxn14345[c0]	Archaeal Lipids	N/A	Manual Addition
ARCSs	Archaeal Lipids	N/A	Manual Addition
CDPDGGR	Archaeal Lipids	N/A	Manual Addition
ASDGGR	Archaeal Lipids	N/A	Manual Addition
rxn10542[c0]	Transport	N/A	Manual Addition
rxn10471[c0]	Transport	N/A	Manual Addition

Table B.2: iMR540 Metabolite Information. A list of every metabolite in the iMR540 reconstruction, including the metabolite ID, name, formula, and charge. Metabolites with the "[c0]" tag are part of the cytosol compartment; metabolites with the "[e0]" tag are part of the extracellular compartment.

Metabolite ID	Metabolite Name	Formula	Charge
cpd02255[c0]	3-oxoadipate-enol-lactone[c0]	C6H5O4	-1
cpd00067[c0]	H[c0]	H	1
cpd00011[c0]	CO2[c0]	CO2	0
cpd00938[c0]	4-Carboxymuconolactone[c0]	C7H4O6	-2
cpd02152[c0]	L-Argininosuccinate[c0]	C10H17N4O6	-1
cpd00106[c0]	Fumarate[c0]	C4H2O4	-2
cpd00051[c0]	L-Arginine[c0]	C6H15N4O2	1
cpd03671[c0]	D-Glucosamine1-phosphate[c0]	C6H14NO8P	0
cpd02611[c0]	N-Acetyl-D-glucosamine1-phosphate[c0]	C8H15NO9P	-1
cpd00010[c0]	CoA[c0]	C21H33N7O16P3S	-3
cpd00022[c0]	Acetyl-CoA[c0]	C23H35N7O17P3S	-3
cpd11420[c0]	trdox[c0]	C6H7NO2R2S2	0
cpd00790[c0]	O-Acetyl-L-homoserine[c0]	C6H11NO4	0
cpd00135[c0]	Homocysteine[c0]	C4H9NO2S	0
cpd00081[c0]	Sulfite[c0]	HO3S	-1
cpd00029[c0]	Acetate[c0]	C2H3O2	-1
cpd00268[c0]	H2S2O3[c0]	S2O3	-2
cpd11421[c0]	trdrd[c0]	C6H9NO2R2S2	0
cpd00060[c0]	L-Methionine[c0]	C5H11NO2S	0
cpd00001[c0]	H2O[c0]	H2O	0
cpd00035[c0]	L-Alanine[c0]	C3H7NO2	0
cpd11590[c0]	met-L-ala-L[c0]	C8H15N2O3S	-1
cpd00161[c0]	L-Threonine[c0]	C4H9NO3	0
cpd11582[c0]	ala-L-Thr-L[c0]	C7H14N2O4	0
cpd03761[c0]	Precorrin_6A[c0]	C44H47N4O16	-7
cpd00006[c0]	NADP[c0]	C21H26N7O17P3	-2
cpd03760[c0]	Precorrin_6B[c0]	C44H49N4O16	-7
cpd00005[c0]	NADPH[c0]	C21H27N7O17P3	-3
cpd00254[e0]	Mg[e0]	Mg	2
cpd00254[c0]	Mg[c0]	Mg	2
cpd00036[c0]	Succinate[c0]	C4H4O4	-2
cpd02857[c0]	DAHP[c0]	C7H11O10P	-2
cpd00699[c0]	5-Dehydroquinate[c0]	C7H9O6	-1
cpd00009[c0]	Phosphate[c0]	HO4P	-2
cpd03519[c0]	2_5-Diaminopyrimidine_nucleoside_triphosphate[c0]	C9H15N5O14P3	-3

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd03518[c0]	Formamidopyrimidine_nucleoside_triphosphate[c0]	C10H15N5O15P3	-3
cpd00047[c0]	Formate[c0]	CHO2	-1
cpd02555[c0]	Tetrahydropteroyltri-L-glutamate[c0]	C29H33N9O12	-4
cpd02738[c0]	5-Methyltetrahydropteroyltri-L-glutamate[c0]	C30H35N9O12	-4
cpd00130[c0]	L-Malate[c0]	C4H4O5	-2
cpd00032[c0]	Oxaloacetate[c0]	C4H2O5	-2
cpd00004[c0]	NADH[c0]	C21H27N7O14P2	-2
cpd00003[c0]	NAD[c0]	C21H26N7O14P2	-1
cpd00346[c0]	L-Aspartate4-semialdehyde[c0]	C4H7NO3	0
cpd00227[c0]	L-Homoserine[c0]	C4H9NO3	0
cpd02656[c0]	6-7-Dimethyl-8-1-D-ribityllumazine[c0]	C13H18N4O6	0
cpd02882[c0]	4-1-D-Ribitylamino-5-aminouracil[c0]	C9H16N4O6	0
cpd00220[c0]	Riboflavin[c0]	C17H20N4O6	0
cpd02893[c0]	5'-Phosphoribosyl-4-carboxy-5-aminoimidazole[c0]	C9H13N3O9P	-1
cpd00002[c0]	ATP[c0]	C10H13N5O13P3	-3
cpd00008[c0]	ADP[c0]	C10H13N5O10P2	-2
cpd02140[c0]	AIR[c0]	C8H14N3O7P	0
cpd00242[c0]	H2CO3[c0]	CHO3	-1
cpd11589[c0]	gly-asp-L[c0]	C6H9N2O5	-1
cpd00041[c0]	L-Aspartate[c0]	C4H6NO4	-1
cpd00033[c0]	Glycine[c0]	C2H5NO2	0
cpd00840[c0]	L-beta-Lysine[c0]	C6H15N2O2	1
cpd00039[c0]	L-Lysine[c0]	C6H15N2O2	1
cpd14960[c0]	Cobalt-precorrin_5B[c0]	C43H44CoN4O16	-6
cpd08371[c0]	Cobalt-precorrin_5[c0]	C45H47CoN4O16	-5
cpd00071[c0]	Acetaldehyde[c0]	C2H4O	0
cpd00123[c0]	3-Methyl-2-oxobutanoate[c0]	C5H7O3	-1
cpd01646[c0]	2-Isopropylmalate[c0]	C7H10O5	-2
cpd00239[c0]	H2S[c0]	H2S	0
cpd15693[c0]	Dianteisopentadecanoylphosphatidylserine[c0]	C36H68NO10P	-2
cpd00054[c0]	L-Serine[c0]	C3H7NO3	0
cpd00046[c0]	CMP[c0]	C9H13N3O8P	-1
cpd15687[c0]	CDP-1_2-dianteisopentadecanoylglycerol[c0]	C42H75N3O15P2	-2
cpd00557[c0]	Siroheme[c0]	C42H36FeN4O16	-8
cpd03426[c0]	Sirohydrochlorin[c0]	C42H38N4O16	-8
cpd10515[c0]	Fe2[c0]	Fe	2
cpd00061[c0]	Phosphoenolpyruvate[c0]	C3H3O6P	-2

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd00482[c0]	2-Phospho-D-glycerate[c0]	C3H5O7P	-2
cpd15682[c0]	1_2-diisohexadecanoyl-sn-glycerol_3-phosphate[c0]	C35H68O8P	-1
cpd15688[c0]	CDP-1_2-diisohexadecanoylglycerol[c0]	C44H79N3O15P2	-2
cpd00012[c0]	PPi[c0]	H2O7P2	-2
cpd00052[c0]	CTP[c0]	C9H13N3O14P3	-3
cpd15421[c0]	CDP-1_2-dioctadecanoylglycerol[c0]	C48H87N3O15P2	-2
cpd15526[c0]	1_2-dioctadecanoyl-sn-glycerol_3-phosphate[c0]	C39H76O8P	-1
cpd02069[c0]	3-Phosphonooxypyruvate[c0]	C3H3O7P	-2
cpd00169[c0]	3-Phosphoglycerate[c0]	C3H5O7P	-2
cpd00038[c0]	GTP[c0]	C10H13N5O14P3	-3
cpd00009[e0]	Phosphate[e0]	HO4P	-2
cpd02333[c0]	Quinolinate[c0]	C7H3NO4	-2
cpd03470[c0]	Iminoaspartate[c0]	C4H3NO4	-2
cpd00095[c0]	Glycerone-phosphate[c0]	C3H6O6P	-1
cpd08366[c0]	2R-Phosphosulfolactate[c0]	C3H4O9PS	-3
cpd02826[c0]	5'-Phosphoribosylformylglycinamidine[c0]	C8H16N3O8P	0
cpd00072[c0]	D-fructose-6-phosphate[c0]	C6H12O9P	-1
cpd00079[c0]	D-glucose-6-phosphate[c0]	C6H12O9P	-1
cpd00118[c0]	Putrescine[c0]	C4H14N2	2
cpd00147[c0]	5-Methylthioadenosine[c0]	C11H15N5O3S	0
cpd00264[c0]	Spermidine[c0]	C7H22N3	3
cpd00837[c0]	S-Adenosylmethioninamine[c0]	C14H24N6O3S	2
cpd00013[c0]	NH3[c0]	NH4	1
cpd00023[c0]	L-Glutamate[c0]	C5H8NO4	-1
cpd00053[c0]	L-Glutamine[c0]	C5H10N2O3	0
cpd00238[c0]	Sedoheptulose7-phosphate[c0]	C7H14O10P	-1
cpd00102[c0]	Glyceraldehyde3-phosphate[c0]	C3H6O6P	-1
cpd00198[c0]	D-Xylulose5-phosphate[c0]	C5H10O8P	-1
cpd00101[c0]	ribose-5-phosphate[c0]	C5H10O8P	-1
cpd00115[c0]	dATP[c0]	C10H13N5O12P3	-3
cpd00246[c0]	Inosine[c0]	C10H12N4O5	0
cpd00114[c0]	IMP[c0]	C10H12N4O8P	-1
cpd00084[c0]	L-Cysteine[c0]	C3H7NO2S	0
cpd15603[c0]	Gly-Cys[c0]	C5H10N2O3S	0
cpd00358[c0]	dUTP[c0]	C9H12N2O14P3	-3
cpd00978[c0]	dUDP[c0]	C9H12N2O11P2	-2
cpd15555[c0]	phosphatidylserine_dihexadecanoyl[c0]	C38H72NO10P	-2

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd15419[c0]	CDP-1_2-dihexadecanoylglycerol[c0]	C44H79N3O15P2	-2
cpd00357[c0]	TTP[c0]	C10H14N2O14P3	-3
cpd00297[c0]	dTDP[c0]	C10H14N2O11P2	-2
cpd00655[c0]	Dephospho-CoA[c0]	C21H33N7O13P2S	-2
cpd00343[c0]	N-Carbamoyl-L-aspartate[c0]	C5H6N2O5	-2
cpd00146[c0]	Carbamoylphosphate[c0]	CH3NO5P	-1
cpd00200[c0]	4MOP[c0]	C6H9O3	-1
cpd00024[c0]	2-Oxoglutarate[c0]	C5H4O5	-2
cpd00107[c0]	L-Leucine[c0]	C6H13NO2	0
cpd01777[c0]	Phosphoribosyl-AMP[c0]	C15H21N5O14P2	-2
cpd01775[c0]	Phosphoribosyl-ATP[c0]	C15H21N5O20P4	-4
cpd00834[c0]	Phosphopantetheine[c0]	C11H22N2O7PS	-1
cpd02666[c0]	R-4'-Phosphopantothenoyl-L-cysteine[c0]	C12H21N2O9PS	-2
cpd02979[c0]	phosphoribosylformiminoaicar-phosphate[c0]	C15H23N5O15P2	-2
cpd02991[c0]	phosphoribulosylformimino-AICAR-phosphate[c0]	C15H23N5O15P2	-2
cpd00290[c0]	D-fructose-1_6-bisphosphate[c0]	C6H12O12P2	-2
cpd00918[c0]	2-Acetamido-5-oxopentanoate[c0]	C7H10NO4	-1
cpd00342[c0]	N-Acetylornithine[c0]	C7H14N2O3	0
cpd11586[c0]	ala-L-glu-L[c0]	C8H13N2O5	-1
cpd00067[e0]	H[e0]	H	1
cpd00205[e0]	K[e0]	K	1
cpd00205[c0]	K[c0]	K	1
cpd00018[c0]	AMP[c0]	C10H13N5O7P	-1
cpd03078[c0]	Selenophosphate[c0]	HO3PSe	-2
cpd01078[c0]	Selenide[c0]	H2Se	0
cpd00091[c0]	UMP[c0]	C9H12N2O9P	-1
cpd00810[c0]	Orotidylic_acid[c0]	C10H11N2O11P	-2
cpd00019[c0]	S-Adenosyl-homocysteine[c0]	C14H20N6O5S	0
cpd01620[c0]	Precorrin_2[c0]	C42H40N4O16	-8
cpd00017[c0]	S-Adenosyl-L-methionine[c0]	C15H23N6O5S	1
cpd03420[c0]	Precorrin_3A[c0]	C43H44N4O16	-6
cpd14961[c0]	Cobalt-precorrin_7[c0]	C45H50CoN4O16	-6
cpd08375[c0]	Cobalt-precorrin_8[c0]	C45H53CoN4O14	-5
cpd00092[c0]	Uracil[c0]	C4H4N2O2	0
cpd00307[c0]	Cytosine[c0]	C4H5N3O	0
cpd15747[c0]	Myristoyllipoteichoic_acid_n=24__linked__unsubstituted[c0]	C115H224O135P24	-24
cpd00014[c0]	UDP[c0]	C9H12N2O12P2	-2

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd15765[c0]	Myristoyllipoteichoic_acid_n=24__linked__N-acetyl-D-glucosamine[c0]	C307H536N24O255P24	-24
cpd00037[c0]	UDP-N-acetylglucosamine[c0]	C17H25N3O17P2	-2
cpd11436[c0]	fa3[c0]	C15H29O2	-1
cpd11437[c0]	fa3coa[c0]	C36H61N7O17P3S	-3
cpd01997[c0]	Dimethylbenzimidazole[c0]	C9H10N2	0
cpd00218[c0]	Niacin[c0]	C6H4NO2	-1
cpd00873[c0]	Nicotinate_ribonucleotide[c0]	C11H13NO9P	-1
cpd02904[c0]	alpha-Ribazole_5'-phosphate[c0]	C14H18N2O7P	-1
cpd03496[c0]	Undecaprenyl-diphospho-N-acetylmuramoyl-N-acetylglucosamine-L-alanyl-D-glutaminyl-meso-2_6-diaminopimeloyl-D-alanyl-D-alanine[c0]	C95H154N9O27P2	-3
cpd03495[c0]	Undecaprenyl-diphospho-N-acetylmuramoyl-N-acetylglucosamine-L-ala-D-glu-meso-2-6-diaminopimeloyl-D-ala-D-ala[c0]	C95H152N8O28P2	-4
cpd11621[c0]	Oxidizedferredoxin[c0]	Fe2R4S6	6
cpd11620[c0]	Reducedferredoxin[c0]	Fe2R4S6	4
cpd08369[c0]	Cobalt-precorrin_3[c0]	C43H42CoN4O16	-6
cpd08368[c0]	Cobalt-precorrin_2[c0]	C42H40CoN4O16	-6
cpd17041[c0]	Protein_biosynthesis[c0]		0
cpd03492[c0]	Undecaprenyl-diphospho-N-acetylmuramoyl-N-acetylglucosamine-L-alanyl-D-isoglutaminyl-L-lysyl-D-alanyl-D-alanine[c0]	C94H155N9O25P2	-2
cpd03491[c0]	Undecaprenyl-diphospho-N-acetylmuramoyl-N-acetylglucosamine-L-alanyl-gamma-D-glutamyl-L-lysyl-D-alanyl-D-alanine[c0]	C94H153N8O26P2	-3
cpd02210[c0]	Indoleglycerol_phosphate[c0]	C11H13NO6P	-1
cpd00359[c0]	indol[c0]	C8H7N	0
cpd02720[c0]	5-Amino-6-5-phosphoribitylaminouracil[c0]	C9H16N4O9P	-1
cpd00931[c0]	5-Amino-6-5-phosphoribosylaminouracil[c0]	C9H14N4O9P	-1
cpd15768[c0]	Anteisoheptadecanoyllipoteichoic_acid_n=24__linked__N-acetyl-D-glucosamine[c0]	C313H548N24O255P24	-24
cpd15750[c0]	Anteisoheptadecanoyllipoteichoic_acid_n=24__linked__unsubstituted[c0]	C121H236O135P24	-24
cpd00113[c0]	Isopentenyldiphosphate[c0]	C5H10O7P2	-2
cpd00289[c0]	Geranylgeranyl_diphosphate[c0]	C20H34O7P2	-2
cpd00350[c0]	Farnesyldiphosphate[c0]	C15H26O7P2	-2
cpd02498[c0]	2_3-Dihydroxy-isovalerate[c0]	C5H9O4	-1
cpd00809[c0]	O-Phospho-L-homoserine[c0]	C4H9NO6P	-1
cpd00361[c0]	ACTN[c0]	C4H8O2	0
cpd00668[c0]	ALCTT[c0]	C5H7O4	-1
cpd00062[c0]	UTP[c0]	C9H12N2O15P3	-3

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd00288[c0]	D-Glucosamine_phosphate[c0]	C6H14NO8P	0
cpd02775[c0]	4-Amino-5-phosphomethyl-2-methylpyrimidine[c0]	C6H9N3O4P	-1
cpd00939[c0]	Toxopyrimidine[c0]	C6H9N3O	0
cpd00868[c0]	p-hydroxyphenylpyruvate[c0]	C9H7O4	-1
cpd00069[c0]	L-Tyrosine[c0]	C9H11NO3	0
cpd00209[c0]	Nitrate[c0]	NO3	-1
cpd00209[e0]	Nitrate[e0]	NO3	-1
cpd03421[c0]	Cobyrinate[c0]	C45H53CoN4O14	-5
cpd17042[c0]	DNA_replication[c0]		0
cpd00149[c0]	Co2[c0]	Co	2
cpd00504[c0]	LL-2_6-Diaminopimelate[c0]	C7H14N2O4	0
cpd00516[c0]	meso-2_6-Diaminopimelate[c0]	C7H14N2O4	0
cpd00132[c0]	L-Asparagine[c0]	C4H8N2O3	0
cpd11581[c0]	gly-asn-L[c0]	C6H11N3O4	0
cpd02978[c0]	7_8-Dihydroneopterin_3'-triphosphate[c0]	C9H13N5O13P3	-3
cpd00177[c0]	dADP[c0]	C10H13N5O9P2	-2
cpd02737[c0]	5-Methyl-H4MPT[c0]	C31H44N6O16P	-3
cpd02438[c0]	Methyl_CoM[c0]	C3H7O3S2	-1
cpd00895[c0]	H4MPT[c0]	C30H42N6O16P	-3
cpd02246[c0]	CoM[c0]	C2H5O3S2	-1
cpd00117[c0]	D-Alanine[c0]	C3H7NO2	0
cpd00128[c0]	Adenine[c0]	C5H5N5	0
cpd02574[c0]	methylthioribose-1-phosphate[c0]	C6H12O7PS	-1
cpd15684[c0]	CDP-1_2-dianteisoheptadecanoylglycerol[c0]	C46H83N3O15P2	-2
cpd15678[c0]	1_2-dianteisoheptadecanoyl-sn-glycerol_3-phosphate[c0]	C37H72O8P	-1
cpd02201[c0]	4-phosphopantothenate[c0]	C9H16NO8P	-2
cpd00508[c0]	3MOP[c0]	C6H9O3	-1
cpd02535[c0]	2_3-Dihydroxy-3-methylvalerate[c0]	C6H11O4	-1
cpd00930[c0]	imidazole_acetol-phosphate[c0]	C6H8N2O5P	-1
cpd00807[c0]	L-histidinol-phosphate[c0]	C6H12N3O4P	0
cpd00015[c0]	FAD[c0]	C27H31N9O15P2	-2
cpd00982[c0]	FADH2[c0]	C27H33N9O15P2	-2
cpd08372[c0]	Cobalt-precorrin_6[c0]	C44H46CoN4O16	-6
cpd00956[c0]	1-2-carboxyphenylamino-1-deoxyribulose_5-phosphate[c0]	C12H14NO9P	-2
cpd00286[c0]	Undecaprenylphosphate[c0]	C55H90O4P	-1
cpd00086[c0]	Propionyl-CoA[c0]	C24H37N7O17P3S	-3

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd00141[c0]	Propionate[c0]	C3H5O2	-1
cpd00236[c0]	D-Erythrose4-phosphate[c0]	C4H8O7P	-1
cpd03706[c0]	UppppU[c0]	C18H22N4O23P4	-4
cpd00274[c0]	Citrulline[c0]	C6H13N3O3	0
cpd00171[c0]	D-Ribulose5-phosphate[c0]	C5H10O8P	-1
cpd15422[c0]	CDP-1_2-ditetradec-7-enoylglycerol[c0]	C40H67N3O15P2	-2
cpd15523[c0]	1_2-ditetradec-7-enoyl-sn-glycerol_3-phosphate[c0]	C31H56O8P	-1
cpd00096[c0]	CDP[c0]	C9H13N3O11P2	-2
cpd01716[c0]	3-Dehydroshikimate[c0]	C7H7O5	-1
cpd08211[c0]	trans_trans_cis-Geranylgeranyl_diphosphate[c0]	C20H34O7P2	-2
cpd02605[c0]	2-isopropyl-3-oxosuccinate[c0]	C7H8O5	-2
cpd02693[c0]	3-Isopropylmalate[c0]	C7H10O5	-2
cpd03608[c0]	4-Hydroxy-L-threonine[c0]	C4H9NO4	0
cpd03607[c0]	4-Phosphonooxy-threonine[c0]	C4H9NO7P	-1
cpd00299[c0]	dUMP[c0]	C9H12N2O8P	-1
cpd00338[c0]	5-Aminolevulinate[c0]	C5H9NO3	0
cpd00689[c0]	Porphobilinogen[c0]	C10H13N2O4	-1
cpd03835[c0]	Precorrin_8[c0]	C45H53N4O14	-7
cpd00065[c0]	L-Tryptophan[c0]	C11H12N2O2	0
cpd00644[c0]	PAN[c0]	C9H16NO5	-1
cpd15554[c0]	phosphatidylserine_ditetradec-7-enoyl[c0]	C34H60NO10P	-2
cpd00863[c0]	beta-D-Glucose_6-phosphate[c0]	C6H12O9P	-1
cpd10162[c0]	R-3-Hydroxy-3-methyl-2-oxopentanoate[c0]	C6H9O4	-1
cpd00533[c0]	dCDP[c0]	C9H13N3O10P2	-2
cpd00356[c0]	dCTP[c0]	C9H13N3O13P3	-3
cpd15604[c0]	Gly-Leu[c0]	C8H16N2O3	0
cpd02884[c0]	FAICAR[c0]	C10H14N4O9P	-1
cpd00175[c0]	UDP-N-acetyl-D-galactosamine[c0]	C17H25N3O17P2	-2
cpd02569[c0]	2-Oxo-3-hydroxyisovalerate[c0]	C5H7O4	-1
cpd00206[c0]	dCMP[c0]	C9H13N3O7P	-1
cpd03834[c0]	Precorrin_4[c0]	C44H45N4O17	-7
cpd03839[c0]	Precorrin_5[c0]	C45H47N4O17	-7
cpd00219[c0]	Prephenate[c0]	C10H8O6	-2
cpd00616[c0]	Pretyrosine[c0]	C10H12NO5	-1
cpd15524[c0]	1_2-dihexadecanoyl-sn-glycerol_3-phosphate[c0]	C35H68O8P	-1
cpd15420[c0]	CDP-1_2-dioctadec-11-enoylglycerol[c0]	C48H83N3O15P2	-2

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd15527[c0]	1_2-dioctadec-11-enoyl-sn-glycerol_3-phosphate[c0]	C39H72O8P	-1
cpd00016[c0]	Pyridoxal_phosphate[c0]	C8H9NO6P	-1
cpd00971[c0]	Na[c0]	Na	1
cpd00971[e0]	Na[e0]	Na	1
cpd00129[c0]	L-Proline[c0]	C5H8NO2	-1
cpd00129[e0]	L-Proline[e0]	C5H8NO2	-1
cpd15557[c0]	phosphatidylserine_dioctadecanoyl[c0]	C42H80NO10P	-2
cpd00203[c0]	1_3-Bisphospho-D-glycerate[c0]	C3H6O10P2	-2
cpd00812[c0]	5-phosphomevalonate[c0]	C6H11O7P	-2
cpd00332[c0]	Mevalonic_acid[c0]	C6H11O4	-1
cpd00143[c0]	Phenylpyruvate[c0]	C9H7O3	-1
cpd03560[c0]	Propionyladenylate[c0]	C13H17N5O8P	-1
cpd00793[c0]	Thiamine_phosphate[c0]	C12H17N4O4PS	0
cpd02894[c0]	4-Amino-2-methyl-5-diphosphomethylpyrimidine[c0]	C6H9N3O7P2	-2
cpd02654[c0]	4-Methyl-5-2-phosphoethyl-thiazole[c0]	C6H9NO4PS	-1
cpd00782[c0]	Pimeloyl-CoA[c0]	C28H42N7O19P3S	-4
cpd01727[c0]	Pimelate[c0]	C7H10O4	-2
cpd11585[c0]	L-alanylglycine[c0]	C5H10N2O3	0
cpd15754[c0]	Isohexadecanoyllipoteichoic_acid_n=24__linked__unsubstituted[c0]	C119H232O135P24	-24
cpd15772[c0]	Isohexadecanoyllipoteichoic_acid_n=24__linked__N-acetyl-D-glucosamine[c0]	C311H544N24O255P24	-24
cpd00103[c0]	PRPP[c0]	C5H10O14P3	-3
cpd15748[c0]	Stearoyllipoteichoic_acid_n=24__linked__unsubstituted[c0]	C123H240O135P24	-24
cpd15766[c0]	Stearoyllipoteichoic_acid_n=24__linked__N-acetyl-D-glucosamine[c0]	C315H552N24O255P24	-24
cpd11593[c0]	ala-L-asp-L[c0]	C7H11N2O5	-1
cpd11440[c0]	fa6[c0]	C16H31O2	-1
cpd11441[c0]	fa6coa[c0]	C37H63N7O17P3S	-3
cpd03666[c0]	2_5-Diamino-6-5'-triphosphoryl-3'_4'-trihydroxy-2'-oxopentyl-_amino-4-oxopyrimidine[c0]	C9H15N5O14P3	-3
cpd00448[c0]	D-Glyceraldehyde[c0]	C3H6O3	0
cpd00100[c0]	Glycerol[c0]	C3H8O3	0
cpd00638[c0]	Deamido-NAD[c0]	C21H24N6O15P2	-2
cpd02851[c0]	AICAR[c0]	C9H14N4O8P	-1
cpd02921[c0]	SAICAR[c0]	C13H16N4O12P	-3
cpd00142[c0]	Acetoacetate[c0]	C4H5O3	-1
cpd00279[c0]	Acetoacetyl-CoA[c0]	C25H37N7O18P3S	-3

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd00020[c0]	Pyruvate[c0]	C3H3O3	-1
cpd00093[c0]	Anthranilate[c0]	C7H6NO2	-1
cpd00216[c0]	Chorismate[c0]	C10H8O6	-2
cpd01017[c0]	Cys-Gly[c0]	C5H10N2O3S	0
cpd00213[c0]	Lipoamide[c0]	C8H15NOS2	0
cpd00449[c0]	Dihydrolipoamide[c0]	C8H17NOS2	0
cpd15692[c0]	Diisopentadecanoylphosphatidylserine[c0]	C36H68NO10P	-2
cpd15686[c0]	CDP-1_2-diisopentadecanoylglycerol[c0]	C42H75N3O15P2	-2
cpd00859[c0]	Pseudouridine_5'-phosphate[c0]	C9H12N2O9P	-1
cpd11580[c0]	Gly-Gln[c0]	C7H13N3O4	0
cpd00025[c0]	H2O2[c0]	H2O2	0
cpd00932[c0]	5-O-1-Carboxyvinyl-3-phosphoshikimate[c0]	C10H10O10P	-3
cpd03049[c0]	2-Hydroxyethyl-ThPP[c0]	C14H21N4O8P2S	-1
cpd00498[c0]	2-Aceto-2-hydroxybutanoate[c0]	C6H9O4	-1
cpd00056[c0]	TPP[c0]	C12H17N4O7P2S	-1
cpd00094[c0]	2-Oxobutyrate[c0]	C4H5O3	-1
cpd00282[c0]	S-Dihydroorotate[c0]	C5H5N2O4	-1
cpd00247[c0]	Orotate[c0]	C5H3N2O4	-1
cpd15685[c0]	CDP-1_2-diisotetradecanoylglycerol[c0]	C40H71N3O15P2	-2
cpd15679[c0]	1_2-diisotetradecanoyl-sn-glycerol_3-phosphate[c0]	C31H60O8P	-1
cpd02817[c0]	HTP[c0]	C11H20NO7PS	-2
cpd02935[c0]	CoM-S-S-CoB[c0]	C13H23NO10PS3	-3
cpd00735[c0]	Formylmethanofuran[c0]	C35H39N4O16	-5
cpd00643[c0]	Methanofuran[c0]	C34H40N4O15	-4
cpd00774[c0]	UroporphyrinogenIII[c0]	C40H36N4O16	-8
cpd00957[c0]	2_5-Diamino-6-5'-phosphoribosylamino-4-pyrimidineone[c0]	C9H15N5O8P	-1
cpd00210[c0]	Taurine[c0]	C2H7NO3S	0
cpd00210[e0]	Taurine[e0]	C2H7NO3S	0
cpd02791[c0]	methylthioribulose-1-phosphate[c0]	C6H12O7PS	-1
cpd17043[c0]	RNA_transcription[c0]		0
cpd00666[c0]	Tartrate[c0]	C4H4O6	-2
cpd02345[c0]	L-Glutamate1-semialdehyde[c0]	C5H9NO3	0
cpd00528[c0]	N2[c0]	N2	0
cpd11640[c0]	H2[c0]	H2	0
cpd00792[c0]	Reduced_coenzyme_F420[c0]	C29H34N5O18P	-4
cpd00649[c0]	Coenzyme_F420[c0]	C29H32N5O18P	-4
cpd00136[c0]	4-Hydroxybenzoate[c0]	C7H5O3	-1

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd02678[c0]	N-Formyl-GAR[c0]	C8H14N2O9P	-1
cpd00492[c0]	N-Acetyl-D-mannosamine[c0]	C8H15NO6	0
cpd11432[c0]	fa11coa[c0]	C38H65N7O17P3S	-3
cpd11431[c0]	fa11[c0]	C17H33O2	-1
cpd00053[e0]	L-Glutamine[e0]	C5H10N2O3	0
cpd00298[c0]	dTMP[c0]	C10H14N2O8P	-1
cpd02030[c0]	3-phosphoshikimate[c0]	C7H9O8P	-2
cpd00322[c0]	L-Isoleucine[c0]	C6H13NO2	0
cpd15553[c0]	phosphatidylserine_ditetradecanoyl[c0]	C34H64NO10P	-2
cpd15423[c0]	CDP-1_2-ditetradecanoylglycerol[c0]	C40H71N3O15P2	-2
cpd00540[c0]	BET[c0]	C5H11NO2	0
cpd00540[e0]	BET[e0]	C5H11NO2	0
cpd00251[c0]	ADPribose[c0]	C15H21N5O14P2	-2
cpd00068[c0]	ITP[c0]	C10H12N4O14P3	-3
cpd00090[c0]	IDP[c0]	C10H12N4O11P2	-2
cpd00151[c0]	Thymine[c0]	C5H6N2O2	0
cpd01587[c0]	5-Methylcytosine[c0]	C5H7N3O	0
cpd15690[c0]	Dianteisoheptadecanoylphosphatidylserine[c0]	C40H76NO10P	-2
cpd00066[c0]	L-Phenylalanine[c0]	C9H11NO2	0
cpd15605[c0]	Gly-Phe[c0]	C11H14N2O3	0
cpd00226[c0]	HYXN[c0]	C5H4N4O	0
cpd00226[e0]	HYXN[e0]	C5H4N4O	0
cpd01710[c0]	2-Isopropylmaleate[c0]	C7H8O4	-2
cpd00202[c0]	DMAPP[c0]	C5H10O7P2	-2
cpd03091[c0]	5'-Deoxyadenosine[c0]	C10H13N5O3	0
cpd01311[c0]	Dethiobiotin[c0]	C10H17N2O3	-1
cpd00104[c0]	BIOT[c0]	C10H15N2O3S	-1
cpd00074[c0]	S[c0]	S	0
cpd02375[c0]	Adenylosuccinate[c0]	C14H15N5O11P	-3
cpd00241[c0]	dGTP[c0]	C10H13N5O13P3	-3
cpd01324[c0]	L-Histidinal[c0]	C6H10N3O	1
cpd00119[c0]	L-Histidine[c0]	C6H9N3O2	0
cpd01080[c0]	ocdca[c0]	C18H35O2	-1
cpd00327[c0]	strcoa[c0]	C39H67N7O17P3S	-3
cpd15522[c0]	1_2-ditetradecanoyl-sn-glycerol_3-phosphate[c0]	C31H60O8P	-1
cpd10515[e0]	Fe2[e0]	Fe	2
cpd00641[c0]	L-Histidinol[c0]	C6H12N3O	1

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd08928[c0]	L-Threonine_phosphate[c0]	C4H9NO6P	-1
cpd02547[c0]	R-1-Aminopropan-2-yl_phosphate[c0]	C3H10NO4P	0
cpd03914[c0]	Cob(II)yrinate_diamide[c0]	C45H57CoN6O12	-3
cpd11584[c0]	Ala-His[c0]	C9H14N4O3	0
cpd11430[c0]	fa1[c0]	C14H27O2	-1
cpd11435[c0]	fa1coa[c0]	C35H59N7O17P3S	-3
cpd08373[c0]	Cobalt-precorrin_6B[c0]	C44H48CoN4O16	-6
cpd00099[c0]	Cl-[c0]	Cl	-1
cpd00099[e0]	Cl-[e0]	Cl	-1
cpd00047[e0]	Formate[e0]	CHO2	-1
cpd11592[c0]	gly-glu-L[c0]	C7H11N2O5	-1
cpd08370[c0]	Cobalt-precorrin_4[c0]	C44H44CoN4O16	-6
cpd01982[c0]	5-Phosphoribosylamine[c0]	C5H12NO7P	0
cpd00755[c0]	Hydroxymethylbilane[c0]	C40H38N4O17	-8
cpd02843[c0]	D-erythro-imidazol-glycerol-phosphate[c0]	C6H10N2O6P	-1
cpd00497[c0]	XMP[c0]	C10H12N4O9P	-1
cpd11587[c0]	Ala-Gln[c0]	C8H15N3O4	0
cpd11225[c0]	3-4-dihydroxy-2-butanone4-phosphate[c0]	C4H8O6P	-1
cpd02679[c0]	5_10-Methylenetetrahydromethanopterin[c0]	C31H42N6O16P	-3
cpd00800[c0]	8-Amino-7-oxononanoate[c0]	C9H17NO3	0
cpd00355[c0]	Nicotinamide_ribonucleotide[c0]	C11H15N2O8P	0
cpd15683[c0]	CDP-1_2-diisoheptadecanoylglycerol[c0]	C46H83N3O15P2	-2
cpd15689[c0]	Diisoheptadecanoylphosphatidylserine[c0]	C40H76NO10P	-2
cpd11438[c0]	fa4[c0]	C15H29O2	-1
cpd11439[c0]	fa4coa[c0]	C36H61N7O17P3S	-3
cpd15680[c0]	1_2-diisopentadecanoyl-sn-glycerol_3-phosphate[c0]	C33H64O8P	-1
cpd00078[c0]	Succinyl-CoA[c0]	C25H36N7O19P3S	-4
cpd15746[c0]	Palmitoyllipoteichoic_acid_n=24__linked__unsubstituted[c0]	C119H232O135P24	-24
cpd15764[c0]	Palmitoyllipoteichoic_acid_n=24__linked__N-acetyl-D-glucosamine[c0]	C311H544N24O255P24	-24
cpd00477[c0]	N-Acetyl-L-glutamate[c0]	C7H9NO5	-2
cpd00026[c0]	UDP-glucose[c0]	C15H22N2O17P2	-2
cpd00144[c0]	UDPglucuronate[c0]	C15H19N2O18P2	-3
cpd02394[c0]	GAR[c0]	C7H15N2O8P	0
cpd00383[c0]	Shikimate[c0]	C7H9O5	-1
cpd00363[c0]	Ethanol[c0]	C2H6O	0
cpd11912[c0]	tRNA-Glu[c0]	C10H12N5O3R	0

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd12227[c0]	L-Glutamyl-tRNA-Glu[c0]	C15H19N6O6R	0
cpd01024[c0]	Methane[c0]	CH4	0
cpd15751[c0]	Isotetradecanoyllipoteichoic_acid_n=24__linked__unsubstituted[c0]	C115H224O135P24	-24
cpd15769[c0]	Isotetradecanoyllipoteichoic_acid_n=24__linked__N-acetyl-D-glucosamine[c0]	C307H536N24O255P24	-24
cpd02642[c0]	N-5-phosphoribosyl-anthranilate[c0]	C12H14NO9P	-2
cpd11434[c0]	fa12coa[c0]	C38H65N7O17P3S	-3
cpd11433[c0]	fa12[c0]	C17H33O2	-1
cpd01695[c0]	Myristoyl-CoA[c0]	C35H59N7O17P3S	-3
cpd03847[c0]	Myristic_acid[c0]	C14H27O2	-1
cpd00307[e0]	Cytosine[e0]	C4H5N3O	0
cpd02552[c0]	n-acetylglutamyl-phosphate[c0]	C7H10NO8P	-2
cpd15417[c0]	CDP-1_2-didodecanoylglycerol[c0]	C36H63N3O15P2	-2
cpd15552[c0]	phosphatidylserine_didodecanoyl[c0]	C30H56NO10P	-2
cpd00822[c0]	O-Succinyl-L-homoserine[c0]	C8H12NO6	-1
cpd03833[c0]	Precorrin_3B[c0]	C43H44N4O17	-6
cpd08210[c0]	ADC[c0]	C10H10NO5	-1
cpd15606[c0]	Gly-Tyr[c0]	C11H14N2O4	0
cpd00334[c0]	L-Lactaldehyde[c0]	C3H6O2	0
cpd00806[c0]	L-Fuculose1-phosphate[c0]	C6H12O8P	-1
cpd00156[c0]	L-Valine[c0]	C5H11NO2	0
cpd00134[c0]	Palmitoyl-CoA[c0]	C37H63N7O17P3S	-3
cpd00214[c0]	Palmitate[c0]	C16H31O2	-1
cpd15767[c0]	Isoheptadecanoyllipoteichoic_acid_n=24__linked__N-acetyl-D-glucosamine[c0]	C313H548N24O255P24	-24
cpd15749[c0]	Isoheptadecanoyllipoteichoic_acid_n=24__linked__unsubstituted[c0]	C121H236O135P24	-24
cpd00031[c0]	GDP[c0]	C10H13N5O11P2	-2
cpd00861[c0]	UDP-N-acetyl-D-mannosamine[c0]	C17H25N3O17P2	-2
cpd00446[c0]	cAMP[c0]	C10H11N5O6P	-1
cpd00182[c0]	Adenosine[c0]	C10H13N5O4	0
cpd00292[c0]	HMG-CoA[c0]	C27H40N7O20P3S	-4
cpd01977[c0]	4-Phospho-L-aspartate[c0]	C4H7NO7P	-1
cpd00283[c0]	Geranyldiphosphate[c0]	C10H18O7P2	-2
cpd03487[c0]	Undecaprenyl-diphospho-N-acetylmuramoyl-N-acetylglucosamine-L-alanyl-D-glutamyl-L-lysyl-D-alanyl-D-alanine[c0]	C94H153N8O26P2	-3
cpd03488[c0]	Undecaprenyl-diphospho-N-acetylmuramoyl-N-acetylglucosamine-L-alanyl-D-glutaminyl-L-lysyl-D-alanyl-D-alanine[c0]	C94H155N9O25P2	-2

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd00089[c0]	Glucose-1-phosphate[c0]	C6H12O9P	-1
cpd15302[c0]	glycogenn-1[c0]	C24H42O21	0
cpd00155[c0]	Glycogen[c0]	C30H52O26	0
cpd03913[c0]	Hydrogenobyrinate_diamide[c0]	C45H58N6O12	-4
cpd03832[c0]	Hydrogenobyrinate[c0]	C45H54N4O14	-6
cpd00764[c0]	7-8-Diaminononanoate[c0]	C9H21N2O2	1
cpd11588[c0]	gly-pro-L[c0]	C7H12N2O3	0
cpd02655[c0]	5_10-Methenyltetrahydromethanopterin[c0]	C31H41N6O16P	-2
cpd00936[c0]	5-Formyl-H4MPT[c0]	C31H42N6O17P	-3
cpd11583[c0]	Ala-Leu[c0]	C9H18N2O3	0
cpd15521[c0]	1_2-didodecanoyl-sn-glycerol_3-phosphate[c0]	C27H52O8P	-1
cpd00043[c0]	UDP-galactose[c0]	C15H22N2O17P2	-2
cpd00092[e0]	Uracil[e0]	C4H4N2O2	0
cpd15691[c0]	Diisotetradecanoylphosphatidylserine[c0]	C34H64NO10P	-2
cpd15269[c0]	octadecenoate[c0]	C18H33O2	-1
cpd15274[c0]	Octadecenoyl-CoA[c0]	C39H65N7O17P3S	-3
cpd11591[c0]	Gly-Met[c0]	C7H14N2O3S	0
cpd15753[c0]	Anteisopentadecanoyllipoteichoic_acid_n=24__linked__unsubstituted[c0]	C117H228O135P24	-24
cpd15771[c0]	Anteisopentadecanoyllipoteichoic_acid_n=24__linked__N-acetyl-D-glucosamine[c0]	C309H540N24O255P24	-24
cpd01914[c0]	L-Methionine_S-oxide[c0]	C5H11NO3S	0
cpd15558[c0]	phosphatidylserine_dioctadec-11-enoyl[c0]	C42H76NO10P	-2
cpd02701[c0]	S-Adenosyl-4-methylthio-2-oxobutanoate[c0]	C15H19N5O6S	0
cpd00149[e0]	Co2[e0]	Co	2
cpd15556[c0]	phosphatidylserine_dihexadec-9-enoyl[c0]	C38H68NO10P	-2
cpd15418[c0]	CDP-1_2-dihexadec-9-enoylglycerol[c0]	C44H75N3O15P2	-2
cpd00521[c0]	dTDP-4-oxo-6-deoxy-D-glucose[c0]	C16H22N2O15P2	-2
cpd02616[c0]	dTDP-4-amino-4_6-dideoxy-D-glucose[c0]	C16H26N3O14P2	-1
cpd12005[c0]	Lipoylprotein[c0]	CHRS2	0
cpd12225[c0]	Dihydrolipolprotein[c0]	CH3RS2	0
cpd15681[c0]	1_2-dianteisopentadecanoyl-sn-glycerol_3-phosphate[c0]	C33H64O8P	-1
cpd15694[c0]	Diisohexadecanoylphosphatidylserine[c0]	C38H72NO10P	-2
cpd00064[c0]	Ornithine[c0]	C5H13N2O2	1
cpd15525[c0]	1_2-dihexadec-9-enoyl-sn-glycerol_3-phosphate[c0]	C35H64O8P	-1
cpd00946[c0]	Undecaprenyl_diphospho_N-acetyl-glucosamine[c0]	C63H103NO12P2	-2
cpd00126[c0]	GMP[c0]	C10H13N5O8P	-1

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd00311[c0]	Guanosine[c0]	C10H13N5O5	0
cpd00485[c0]	D-Mannose1-phosphate[c0]	C6H12O9P	-1
cpd00235[c0]	D-mannose-6-phosphate[c0]	C6H12O9P	-1
cpd00305[c0]	Thiamin[c0]	C12H17N4OS	1
cpd00305[e0]	Thiamin[e0]	C12H17N4OS	1
cpd00295[c0]	dGDP[c0]	C10H13N5O10P2	-2
cpd00073[c0]	Urea[c0]	CH4N2O	0
cpd00073[e0]	Urea[e0]	CH4N2O	0
cpd15238[c0]	Hexadecenoyl-CoA[c0]	C37H61N7O17P3S	-3
cpd15237[c0]	hexadecenoate[c0]	C16H29O2	-1
cpd00830[c0]	4-Hydroxy-2-oxoglutarate[c0]	C5H4O6	-2
cpd01974[c0]	4-Hydroxy-L-glutamate[c0]	C5H8NO5	-1
cpd00712[c0]	2-Dehydropantoate[c0]	C6H9O4	-1
cpd00408[c0]	Pantoate[c0]	C6H11O4	-1
cpd02636[c0]	4-Methyl-5-2-hydroxyethyl-thiazole[c0]	C6H9NOS	0
cpd00367[c0]	Cytidine[c0]	C9H13N3O5	0
cpd00738[c0]	phosphoserine[c0]	C3H7NO6P	-1
cpd15770[c0]	Isopentadecanoyllipoteichoic_acid_n=24__linked__N-acetyl-D-glucosamine[c0]	C309H540N24O255P24	-24
cpd15752[c0]	Isopentadecanoyllipoteichoic_acid_n=24__linked__unsubstituted[c0]	C117H228O135P24	-24
cpd15677[c0]	1_2-diisoheptadecanoyl-sn-glycerol_3-phosphate[c0]	C37H72O8P	-1
cpd00152[c0]	Agmatine[c0]	C5H16N4	2
cpd10516[c0]	fe3[c0]	Fe	3
cpd10516[e0]	fe3[e0]	Fe	3
cpd00001[e0]	H2O[e0]	H2O	0
cpd00011[e0]	CO2[e0]	CO2	0
cpd02465[c0]	tetrahydrodipicolinate[c0]	C7H7NO4	-2
cpd02211[c0]	L-2-Amino-acetoacetate[c0]	C4H7NO3	0
cpd00058[c0]	Cu2[c0]	Cu	2
cpd00042[c0]	GSH[c0]	C10H16N3O6S	-1
cpd12370[c0]	apo-ACP[c0]	HOR	0
cpd11416[c0]	Biomass[c0]		0
cpd00063[c0]	Ca2[c0]	Ca	2
cpd03422[c0]	Cobinamide[c0]	C48H73CoN11O8	3
cpd00166[c0]	Calomide[c0]	C72H101CoN18O17P	1
cpd11493[c0]	ACP[c0]	C11H21N2O7PRS	-1
cpd03443[c0]	3-Octaprenyl-4-hydroxybenzoate[c0]	C47H69O3	-1

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd03444[c0]	2-Octaprenylphenol[c0]	C46H70O	0
cpd11524[c0]	5-methyl-hexanoyl-ACP[c0]	C18H33N2O8PRS	-1
cpd01772[c0]	Succinylbenzoate[c0]	C11H8O5	-2
cpd03451[c0]	SHCHC[c0]	C11H10O6	-2
cpd00421[c0]	Triphosphate[c0]	HO10P3	-4
cpd11492[c0]	Malonyl-acyl-carrierprotein-[c0]	C14H22N2O10PRS	-2
cpd11525[c0]	7-methyl-3-oxo-octanoyl-ACP[c0]	C20H35N2O9PRS	-1
cpd00034[e0]	Zn2[e0]	Zn	2
cpd11496[c0]	4-methyl-3-oxo-hexanoyl-ACP[c0]	C18H31N2O9PRS	-1
cpd11495[c0]	2-methylbutyryl-ACP[c0]	C16H29N2O8PRS	-1
cpd02083[c0]	CoproporphyrinogenIII[c0]	C36H40N4O8	-4
cpd00817[c0]	D-Arabinose5-phosphate[c0]	C5H10O8P	-1
cpd11515[c0]	12-methyl-tetra-decanoyl-ACP[c0]	C26H49N2O8PRS	-1
cpd01741[e0]	ddca[e0]	C12H23O2	-1
cpd11488[c0]	Acetoacetyl-ACP[c0]	C15H25N2O9PRS	-1
cpd15268[c0]	Octadecanoyl-ACP[c0]	C29H55N2O8PRS	-1
cpd01270[c0]	FMNH2[c0]	C17H22N4O9P	-1
cpd04122[c0]	Aminoacetaldehyde[c0]	C2H6NO	1
cpd00050[c0]	FMN[c0]	C17H20N4O9P	-1
cpd00027[c0]	D-Glucose[c0]	C6H12O6	0
cpd11532[c0]	9-methyl-decanoyl-ACP[c0]	C22H41N2O8PRS	-1
cpd00080[c0]	Glycerol-3-phosphate[c0]	C3H8O6P	-1
cpd11533[c0]	11-methyl-3-oxo-dodecanoyl-ACP[c0]	C24H43N2O9PRS	-1
cpd03918[c0]	Adenosyl_cobinamide[c0]	C58H85CoN16O11	2
cpd00355[e0]	Nicotinamide_ribonucleotide[e0]	C11H15N2O8P	0
cpd00626[c0]	dTDPglucose[c0]	C16H24N2O16P2	-2
cpd02120[c0]	Dihydrodipicolinate[c0]	C7H5NO4	-2
cpd03448[c0]	2-Octaprenyl-3-methyl-6-methoxy-1_4-benzoquinone[c0]	C48H72O3	0
cpd11516[c0]	14-methyl-3-oxo-hexa-decanoyl-ACP[c0]	C28H51N2O9PRS	-1
cpd11507[c0]	8-methyl-decanoyl-ACP[c0]	C22H41N2O8PRS	-1
cpd11508[c0]	10-methyl-3-oxo-dodecanoyl-ACP[c0]	C24H43N2O9PRS	-1
cpd11504[c0]	8-methyl-3-oxo-decanoyl-ACP[c0]	C22H39N2O9PRS	-1
cpd03447[c0]	2-Octaprenyl-6-methoxy-1_4-benzoquinone[c0]	C47H70O3	0
cpd00058[e0]	Cu2[e0]	Cu	2
cpd00906[c0]	all-trans-Hexaprenyl_diphosphate[c0]	C30H50O7P2	-2
cpd11540[c0]	13-methyl-tetra-decanoyl-ACP[c0]	C26H49N2O8PRS	-1
cpd00760[c0]	2-Methylbutyryl-CoA[c0]	C26H41N7O17P3S	-3

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd11499[c0]	4-methyl-hexanoyl-ACP[c0]	C18H33N2O8PRS	-1
cpd03919[c0]	Adenosyl_cobinamide_phosphate[c0]	C58H84CoN16O14P	0
cpd03917[c0]	Adenosylcobyric_acid[c0]	C55H77CoN15O11	1
cpd02039[c0]	1-Aminopropan-2-ol[c0]	C3H10NO	1
cpd02590[c0]	all-trans-Heptaprenyl_diphosphate[c0]	C35H58O7P2	-2
cpd02557[c0]	Farnesylfarnesylgeraniol[c0]	C40H66O7P2	-2
cpd11484[c0]	HMA[c0]	C25H47N2O9PRS	-1
cpd11529[c0]	9-methyl-3-oxo-decanoyl-ACP[c0]	C22H39N2O9PRS	-1
cpd00558[e0]	Spermine[e0]	C10H30N4	4
cpd15479[c0]	glucosyl-inner_core_oligosaccharide_lipid_A[c0]	C151H265N2O79P4	-7
cpd00030[e0]	Mn2[e0]	Mn	2
cpd02685[c0]	N-acetyl-LL-2_6-diaminopimelate[c0]	C9H15N2O5	-1
cpd03916[c0]	Adenosyl_cobyrinate_diamide[c0]	C55H69CoN11O15	-3
cpd03915[c0]	Cob(I)yrinate_diamide[c0]	C45H57CoN6O12	-4
cpd00063[e0]	Ca2[e0]	Ca	2
cpd03920[c0]	Adenosylcobinamide-GDP[c0]	C68H96CoN21O21P2	0
cpd11503[c0]	6-methyl-octanoyl-ACP[c0]	C20H37N2O8PRS	-1
cpd11536[c0]	11-methyl-dodecanoyl-ACP[c0]	C24H45N2O8PRS	-1
cpd00655[e0]	Dephospho-CoA[e0]	C21H33N7O13P2S	-2
cpd03289[c0]	L-2-Acetamido-6-oxopimelate[c0]	C9H11NO6	-2
cpd08316[c0]	D-Glycero-D-manno-heptose1-7-bisphosphate[c0]	C7H14O13P2	-2
cpd15489[c0]	inner_core_oligosaccharide_lipid_A[c0]	C145H255N2O74P4	-7
cpd04920[c0]	D-Glycero-D-manno-heptose1-phosphate[c0]	C7H14O10P	-1
cpd11521[c0]	5-methyl-3-oxo-hexanoyl-ACP[c0]	C18H31N2O9PRS	-1
cpd11520[c0]	isovaleryl-ACP[c0]	C16H29N2O8PRS	-1
cpd02021[c0]	Succinylbenzoyl-CoA[c0]	C32H40N7O20P3S	-4
cpd11511[c0]	10-methyl-dodecanoyl-ACP[c0]	C24H45N2O8PRS	-1
cpd00045[c0]	Adenosine_3-5-bisphosphate[c0]	C10H13N5O10P2	-2
cpd02886[c0]	UDP-3-O-beta-hydroxymyristoyl-N-acetylglucosamine[c0]	C31H51N3O19P2	-2
cpd03423[c0]	alpha-Ribazole[c0]	C14H18N2O4	0
cpd11537[c0]	13-methyl-3-oxo-tetra-decanoyl-ACP[c0]	C26H47N2O9PRS	-1
cpd15269[e0]	octadecenoate[e0]	C18H33O2	-1
cpd11541[c0]	15-methyl-3-oxo-hexa-decanoyl-ACP[c0]	C28H51N2O9PRS	-1
cpd11528[c0]	7-methyl-octanoyl-ACP[c0]	C20H37N2O8PRS	-1
cpd03494[c0]	Undecaprenyl-diphospho-N-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2-6-diaminopimeloyl-D-alanyl-D-alanine[c0]	C87H139N7O23P2	-4

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd15358[c0]	2-octadecanoyl-sn-glycerol_3-phosphate[c0]	C21H42O7P	-1
cpd02295[c0]	1-4-Dihydroxy-2-naphthoate[c0]	C11H7O4	-1
cpd11512[c0]	12-methyl-3-oxo-tetra-decanoyl-ACP[c0]	C26H47N2O9PRS	-1
cpd00111[c0]	Oxidized_glutathione[c0]	C20H30N6O12S2	-2
cpd03422[e0]	Cobinamide[e0]	C48H73CoN11O8	3
cpd02968[c0]	UDP-N-acetylmuramoyl-L-alanyl-D-glutamyl-6-carboxy-L-lysyl-D-alanyl-_D-alanine[c0]	C41H61N9O28P2	-4
cpd00658[c0]	Isochorismate[c0]	C10H8O6	-2
cpd01080[e0]	ocdca[e0]	C18H35O2	-1
cpd00111[e0]	Oxidized_glutathione[e0]	C20H30N6O12S2	-2
cpd11500[c0]	6-methyl-3-oxo-octanoyl-ACP[c0]	C20H35N2O9PRS	-1
cpd03847[e0]	Myristic_acid[e0]	C14H27O2	-1
cpd00869[c0]	4-methylthio_2-oxobutyrate[c0]	C5H7O3S	-1
cpd11295[c0]	2_3-diketo5-methylthio-1-phosphopentane[c0]	C6H10O6PS	-1
cpd11217[c0]	1_4-Dihydroxy-2-naphthoyl-CoA[c0]	C32H39N7O19P3S	-3
cpd00070[c0]	Malonyl-CoA[c0]	C24H34N7O19P3S	-4
cpd00085[c0]	beta-Alanine[c0]	C3H7NO2	0
cpd00506[c0]	gamma-Glutamylcysteine[c0]	C8H13N2O5S	-1
cpd00013[e0]	NH3[e0]	NH4	1
cpd02591[c0]	pendp[c0]	C25H42O7P2	-2
cpd16335[c0]	2-Succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate[c0]	C14H13O9	-3
cpd00460[c0]	sulfoacetaldehyde[c0]	C2H3O4S	-1
cpd03285[c0]	3-sulfopyruvate[c0]	C3H2O6S	-2
cpd08367[c0]	(2R)-3-sulfolactate[c0]	C3H4O6S	-2
cpd00607[c0]	Citramalate[c0]	C5H6O5	-2
cpd01502[c0]	Citraconate[c0]	C5H4O4	-2
cpd03593[c0]	D-erythro-3-methylmalate[c0]	C5H6O5	0
cpd00029[e0]	Acetate[e0]	C2H3O2	-1
cpd00278[c0]	Indole-3-pyruvate[c0]	C11H8NO3	-1
cpd11175[c0]	S-2-(indol-3-yl)acetyl-CoA[c0]	C31H40N8O17P3S	-3
cpd00035[e0]	L-Alanine[e0]	C3H7NO2	0
cpd00117[e0]	D-Alanine[e0]	C3H7NO2	0
cpd00528[e0]	N2[e0]	N2	0
cpd00239[e0]	H2S[e0]	H2S	0
cpd15886[c0]	trans-homoaconitate[c0]	C7H5O6	-3
cpd15833[c0]	S-homocitrate[c0]	C7H7O7	-3
cpd02483[c0]	cis-Homoaconitate[c0]	C7H5O6	-3
cpd15888[c0]	threo-isohomocitrate[c0]	C7H7O7	-3

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd15901[c0]	2-oxohexanedioic_acid[c0]	C6H6O5	-2
cpd15831[c0]	(R)-(homo)2citrate[c0]	C8H9O7	-3
cpd15882[c0]	cis-(homo)2aconitate[c0]	C8H7O6	-3
cpd15908[c0]	(-)threo-iso(homo)2citrate[c0]	C8H9O7	-3
cpd15900[c0]	2-oxoheptanedioic_acid[c0]	C7H8O5	-2
cpd15832[c0]	(R)-(homo)3citrate[c0]	C9H11O7	-3
cpd15883[c0]	cis-(homo)3aconitate[c0]	C9H9O6	-3
cpd15909[c0]	(-)threo-iso(homo)3citrate[c0]	C9H11O7	-3
cpd16398[c0]	2-Oxosuberate[c0]	C8H10O5	-2
cpd15829[c0]	7-oxoheptanoic_acid[c0]	C7H11O3	-1
cpd15827[c0]	7-mercaptoheptanoic_acid[c0]	C7H13O2S	-1
cpd15828[c0]	7-mercaptoheptanoylthreonine[c0]	C11H20NO4S	-1
cpd15850[c0]	7,8-dihydronepterin_2_3-cyclicphosphate[c0]	C9H11N5O6P	-1
cpd03521[c0]	Dihydroneopterin_phosphate[c0]	C9H13N5O7P	-1
cpd02961[c0]	Dihydroneopterin[c0]	C9H13N5O4	0
cpd00954[c0]	6-hydroxymethyl_dihydropterin[c0]	C7H9N5O2	0
cpd00229[c0]	Glycolaldehyde[c0]	C2H4O2	0
cpd02920[c0]	2-Amino-4-hydroxy-6-hydroxymethyl-7-8-dihydropteridinediphosphate[c0]	C7H9N5O8P2	-2
cpd00443[c0]	ABEE[c0]	C7H6NO2	-1
cpd15830[c0]	4-(B-D-ribofuranosyl)aminobenzene_5-phosphate[c0]	C11H14NO7P	-2
cpd15851[c0]	7,8-dihydropterin-6-ylmethyl-4-(B-D-ribofuranosyl)_aminobenzene_5-phosphate[c0]	C18H21N6O8P	-2
cpd02041[c0]	(S)-2-Hydroxyglutarate[c0]	C5H6O5	-2
cpd15853[c0]	6-deoxy-5-ketofructose-1-phosphate[c0]	C6H9O8P	-2
2ATDLH6U[c0]	2-amino-2,3,7-trideoxy-D-lyxo-hept-6-ulosonate[c0]	C7H11NO5	-2
cpd17158[c0]	Hydroxypyruvaldehyde_phosphate[c0]	C3H5O6P	0
2A3DHQ[c0]	4-Amino-3-Dehydroquinate[c0]	C7H10O5N	-1
4A3DHS[c0]	4-Amino-3-Dehydroshikimate[c0]	C7H8O4N	-1
4ASKM[c0]	4-Aminoshikimate[c0]	C7H10O4N	-1
4A3H15D1C[c0]	4-amino-3-hydroxycyclohexa-1,5-diene-1-carboxylate[c0]	C7H8O3N	-1
cpd00139[c0]	Glycolate[c0]	C2H3O3	-1
cpd00040[c0]	Glyoxalate[c0]	C2HO3	-1
cpd00374[c0]	Tyramine[c0]	C8H12NO	1
GGT[c0]	gamma-Glutamyl-tyramine[c0]	C13H18N2O4	0
4HM2FCP[c0]	4-(hydroxymethyl)-2-furancarboxaldehyde-phosphate[c0]	C6H6O6P	-1

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
5AM3FMP[c0]	5-(aminomethyl)-3-furanmethanol-phosphate[c0]	C6H10O5NP	0
5AM3FMPP[c0]	5-(aminomethyl)-3-furanmethanol-pyrophosphate[c0]	C6H10O8NP2	-1
AEPM2FMA[c0]	4((4-(2-aminoethyl)phenoxy)methyl)-2-furanmethanamine[c0]	C19H26O5N3	1
cpd00244[c0]	Ni2[c0]	Ni	2
cpd15873[c0]	Pyrrochorphinate[c0]	C42H42N6NiO14	-6
cpd15874[c0]	Dihydrocorphinate[c0]	C42H45N6NiO14	-5
cpd15875[c0]	Tetrahydrocorphinate[c0]	C42H47N6NiO14	-5
cpd15905[c0]	15_17-seco-F430-17-acid[c0]	C42H47N6NiO14	-5
cpd03425[c0]	Factor_430[c0]	C42H46N6O13Ni	-4
cpd00244[e0]	Ni2[e0]	Ni	2
cpd00180[c0]	Oxalate[c0]	C2O4	-2
cpd15839[c0]	7,8-didemethyl-8-hydroxy-5-deazariboflavin[c0]	C16H17N3O7	0
cpd00159[c0]	L-Lactate[c0]	C3H5O3	-1
cpd15809[c0]	2-phospho-L-lactate[c0]	C3H4O6P	-3
cpd15889[c0]	lactyl-(2)-diphospho-(5)-guanosine[c0]	C13H16N5O13P2	-3
cpd15864[c0]	Coenzyme_F420-0[c0]	C19H20N3O12P	-2
cpd15865[c0]	Coenzyme_F420-1[c0]	C24H26N4O15P	-3
cpd15868[c0]	Coenzyme_F420-3[c0]	C34H38N6O21P	-5
cpd00204[c0]	CO[c0]	CO	1
cpd00204[e0]	CO[e0]	CO	1
cpd00131[e0]	Molybdenum[e0]	Mo	0
cpd00131[c0]	Molybdenum[c0]	Mo	0
cpd03523[c0]	7,8-Dihydromethanopterin[c0]	C30H40N6O16P	-3
cpd03732[c0]	UDP-N-acetyl-D-mannosaminouronate[c0]	C17H22N3O18P2	-3
cpd02782[c0]	UDP-N-acetyl-D-glucosaminouronate[c0]	C17H22N3O18P2	-3
U2A2D3OG[c0]	UDP-2-acetamido-2-deoxy-3-oxo-glucuronate[c0]	C17H20N3O18P2	-2
U2A3A23DDG[c0]	UDP-2-acetamido-3-amino-2,3-dideoxy-glucuronate[c0]	C17H23N4O17P2	-2
U23DA23DDG[c0]	UDP-2,3-diacetamido-2,3-dideoxy-glucuronate[c0]	C19H25N4O18P2	-2
U23DA23DDM[c0]	UDP-2,3-diacetamido-2,3-dideoxy-mannuronate[c0]	C19H25N4O18P2	-2
U3A23DAM[c0]	UDP-3-acetamido-2,3-diaminomannuronate[c0]		0
N2A24D5MH4U15P[c0]	NDP-2-acetamino-2,4-dideoxy-5-O-methyl-hexos-5-ulo-1,5-pyranose[c0]		0
N2A24D5MAEH5U15P[c0]	NDP-(5S)-2-acetamido-2,4-dideoxy-5-O-methyl-alpha-L-erythro-hexos-5-ulo-1,5-pyranose[c0]		0
LIP4SUG[c0]	Lipid_tetrasaccharide[c0]		0

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
LIP4SUGT[c0]	Lipid_tetrasaccharide_Thr[c0]		0
LIP1SUG[c0]	Lipid_monosaccharide[c0]		0
LIP2SUG[c0]	Lipid_disaccharide[c0]		0
LIP3SUG[c0]	Lipid_trisaccharide[c0]		0
LIP4SUGT[e0]	Lipid_tetrasaccharide_Thr[e0]		0
FLGN[e0]	Flagellin[e0]		0
ARCN[e0]	Archaellin[e0]		0
MEMLIP[c0]	Membrane_lipid[c0]		0
cpd00703[c0]	Indoleacetate[c0]	C10H8NO2	-1
cpd00703[e0]	Indoleacetate[e0]	C10H8NO2	-1
cpd00489[e0]	4-Hydroxyphenylacetate[e0]	C8H7O3	-1
cpd00489[c0]	4-Hydroxyphenylacetate[c0]	C8H7O3	-1
cpd00430[e0]	PACT[e0]	C8H8O2	0
cpd00430[c0]	PACT[c0]	C8H8O2	0
cpd03165[c0]	4-Hydroxyphenylacetyl-CoA[c0]	C29H39N7O18P3S	-3
cpd00452[c0]	Phenylacetyl-CoA[c0]	C29H39N7O17P3S	-3
cpd00802[c0]	D-fructose-1-phosphate[c0]	C3H4O2	0
cpd00428[c0]	2-Oxopropanal[c0]	C3H4O2	0
cpd00055[c0]	Formaldehyde[c0]	CH2O	0
cpd15573[c0]	tRNA(SeCys)[c0]	C10H12N5O3R	0
cpd15565[c0]	L-Seryl-tRNA(Sec)[c0]	C18H28NO18P2R3	-2
cpd16442[c0]	O-Phosphoseryl-tRNA(Sec)[c0]	C12H15N3O5S	0
cpd15563[c0]	L-Selenocysteinyl-tRNA(Sec)[c0]	C18H30NO17P2R3Se	0
cpd16579[c0]	Selenoprotein[c0]	C5H12O	0
cpd03387[c0]	Selenite[c0]	H2O3Se	2
cpd03396[c0]	Selenate[c0]	H2O4Se	4
cpd03396[e0]	Selenate[e0]	H2O4Se	4
cpd00207[c0]	Guanine[c0]	C5H5N5O	0
cpd17039[c0]	Isopentenylphosphate[c0]	C5H9O4P	-2
cpd02797[c0]	sn-3-O-(Geranylgeranyl)glycerol_1-phosphate[c0]	C23H40O6P	-1
cpd02824[c0]	2,3-Bis-O-(geranylgeranyl)glycerol_1-phosphate[c0]	C43H72O6P	-1
cpd18042[c0]	CDP-2,3-bis-O-(geranylgeranyl)-sn-glycerol[c0]	C52H83N3O13P2	-2
ARCHLS[c0]	Archaetidylserine[c0]	C46H77NO8P	-1
SATARCHL[c0]	Saturated_CDP-archaeol[c0]	C52H99N3O13P2	-2
SATARCHLS[c0]	Saturated_Archaetidylserine[c0]	C46H93NO8P	-1
cpd11640[e0]	H2[e0]	H2	0

Table B.2 (cont.)

Metabolite ID	Metabolite Name	Formula	Charge
cpd01024[e0]	Methane[e0]	CH4	0
SATARCHLS[c0]	Saturated_Archaetidylserine[c0]	C46H93NO8P	-1
cpd11640[e0]	H2[e0]	H2	0
cpd01024[e0]	Methane[e0]	CH4	0

Table B.3: McNA Chemostat Medium. This is the recipe for the medium used in chemostat experiments described in Chapter 3 Methods and simulated by the "maxGrowthOnH2" script (see Table B.4). Importantly, this recipe is for liquid media only and does not contain gases; these are listed in Chapter 3 methods.

Compound	Amount (g)	Amount (mL)	Final Concentration (mM)
H_2O	-	8289	-
KCl	3.02	-	4.5
$NaHCO_3$	45	-	59.5
NaCl	198	-	376.5
$NaC_2H_3O_2{\cdot}3H_2O$	12.6	-	10
$FeSO_4$ Solution	-	45	-
Resazurin Solution	-	9	-
Trace Minerals Solution	-	9	-
Vitamin Solution	-	90	-
Cysteine HCl	4.5	-	2.85
Divalent Cation Solution	-	500	-
NH_4Cl Solution	-	18	-
K_2HPO_4 Solution	-	90	-
$FeSO_4$ Solution (per 100 mL of 10 mM HCl)			
$FeSO_4{\cdot}7H_2O$	0.19	-	0.034
Resazurin Solution (per L H_2O)			
Resazurin	1	-	0.004
1000X Trace Mineral Solution (per 100 mL H_2O)			
$Na_3Citrate{\cdot}2H_2O$	2.1	-	0.081
$MnSO_4{\cdot}H_2O$	0.5	-	0.030
$CoCl_2{\cdot}6H_2O$	0.1	-	0.004
$ZnSO_4{\cdot}7H_2O$	0.1	-	0.003
$CuSO_4{\cdot}5H_2O$	0.01	-	4.00E-04
$AlK(SO_4)_2$	0.01	-	3.87E-04
H_3BO_4	0.01	-	0.001
$Na_2MoO_4{\cdot}2H_2O$	0.1	-	0.002
$NiCl_2{\cdot}6H_2O$	0.025	-	0.001
Na_2SeO_3	0.2	-	0.012
VCl_3	0.01	-	6.36E-04

Table B.3 (cont.)

Compound	Amount (g)	Amount (mL)	Final Concentration (mM)
$Na_2WO_4 \cdot 2H_2O$	0.0033	-	1.00E-04
100X Vitamin Solution (per L H_2O)			
Biotin	0.002	-	8.19E-05
Folic Acid	0.002	-	4.53E-05
Pyridoxine HCl	0.01	-	4.86E-04
Thiamine HCl	0.005	-	1.48E-04
Riboflavin	0.005	-	1.33E-04
Nicotinic Acid	0.005	-	4.06E-04
DL-Calcium Pantothenate	0.005	-	1.05E-04
Vitamin B_{12}	0.0001	-	7.38E-07
p-Aminobenzoic Acid	0.005	-	3.65E-04
Lipoic Acid	0.005	-	2.42E-04
5 M NH4Cl Solution (per L H_2O)			
NH_4Cl	267.5	-	10
80 mM K_2HPO_4 Solution (per L H_2O)			
K_2HPO_4	14	-	0.8
Divalent Cation Solution (per 500 mL H_2O)			
$CaCl_2 \cdot 2H_2O$	1.26	-	8.6
$MgCl_2 \cdot 6H_2O$	24.75	-	121.7
$MgSO_4 \cdot 7H_2O$	31.05	-	126.0

Table B.4: iMR540 *in silico* Medium. This is the medium for the "maxGrowthOnH2.m" script, which simulates growth of *M. maripaludis* on H_2+CO_2 with acetate supplementation. It contains some, but not all, of the trace minerals in McNA chemostat medium (see Table B.3). Adding these compounds as possible *in silico* medium components would not affect simulations because the model does not require these compounds to simulate growth, thus they are not included here.

Compound ID	Compound Name	Reaction ID
cpd00001	H2O	EX_cpd00001[e0]
cpd00009	Phosphate	EX_cpd00009[e0]
cpd00011	CO2	EX_cpd00011[e0]
cpd00013	NH3	EX_cpd00013[e0]
cpd00029	Acetate	EX_cpd00029[e0]
cpd00030	Mn2+	EX_cpd00030[e0]
cpd00034	Zn2+	EX_cpd00034[e0]
cpd00058	Cu2+	EX_cpd00058[e0]
cpd00063	Ca2+	EX_cpd00063[e0]
cpd00099	Cl-	EX_cpd00099[e0]
cpd00131	Molybdenum	EX_cpd00131[e0]
cpd00149	Co2+	EX_cpd00149[e0]
cpd00205	K+	EX_cpd00205[e0]
cpd00239	H2S	EX_cpd00239[e0]
cpd00244	Ni2+	EX_cpd00244[e0]
cpd00254	Mg	EX_cpd00254[e0]
cpd00305	Thiamin	EX_cpd00305[e0]
cpd00355	Nicotinamide ribonucleotide	EX_cpd00355[e0]
cpd03396	Selenate	EX_cpd03396[e0]
cpd10515	Fe2+	EX_cpd10515[e0]
cpd10516	Fe3	EX_cpd10516[e0]
cpd11640	H2	EX_cpd11640[e0]
cpd15269	Octadecenoate	EX_cpd15269[e0]

Supplementary Information B.1: Comparison with Genome-Scale Essentiality Indices. The following short analysis compares gene knockout predictions from iMR540 against essentiality index predictions, another model that attempts to evaluate essential genes.

A previous group performed a genome-scale analysis gene function of *Methanococcus maripaludis* via a saturated mutagenesis technique on rich and minimal media [163]. Although this dataset does not contain the same quality of knockout data as actual knockout experiments, it provides a valuable "first pass" test set for gene essentiality of our model. For minimal medium in particular, their data included 2 whole genome libraries of mapped insertions, each of which contained growth data for 7 (T1) and 14 generations (T2). Reasoning that essential genes would likely be conserved across mutants, they correlated number of insertions at a particular gene location with gene essentiality by calculating an "essentiality index" (EI) for each location. Based upon a set of "known essential" genes, they set a cutoff of EI ≤ 3 for essential genes, effectively creating predictions of gene essentiality for all genes.

Considering the 4 sets of library:generation combinations—Lib.1:T1, Lib.1:T2, Lib.2:T1, Lib.2:T2—each gene could be predicted to be essential in 0-4 cases. Rather than globally classify gene essentiality based on all 4 cases, we created 4 separate sets of essential genes by setting different essentiality thresholds. For example, in "4 instances", only genes that were predicted as essential in all 4 libraries were treated as essential genes and all other genes were considered non-essential; in "1 instance", all genes that were predicted as essential in at least 1 library were treated as essential genes. The iMR540 reconstruction shared 538 genes with this dataset, thus we were able to compare gene essentiality predictions across nearly the entire model.

As shown by Figure B.1, different thresholds had a great effect on the EI predictions; a lower threshold necessarily caused an increase in negative (no-growth) outcomes and a decrease in

positive (yes-growth) outcomes. Our model experienced no change in its gene essentiality predictions in relation to threshold, hence a decrease in threshold resulted in improved performance on negative predictions and decreased performance on positive predictions. The threshold's effect on overall performance, displayed in Figure B.2, shows that our model's predictive accuracy in the four cases ranged from 61.3-65.2% and was maximized in the "3 instances" dataset, whereas MCC ranged from 0.283-0.326 and was highest for "2 instances". This small discrepancy reflects the difference in how these metrics are calculated, with MCC putting greater emphasis on our model's improved ability to predict true negative outcomes.

Overall, this analysis revealed a slight positive correlation between EI predictions and gene essentiality predictions from out model. It is important to keep in mind that EI, like our reconstruction, is a model of gene essentiality and should not be confused for actual knockout data. Through different methods, both models provide hypotheses for gene functions outside known metabolism and could fuel future investigations to directly measure gene essentiality.

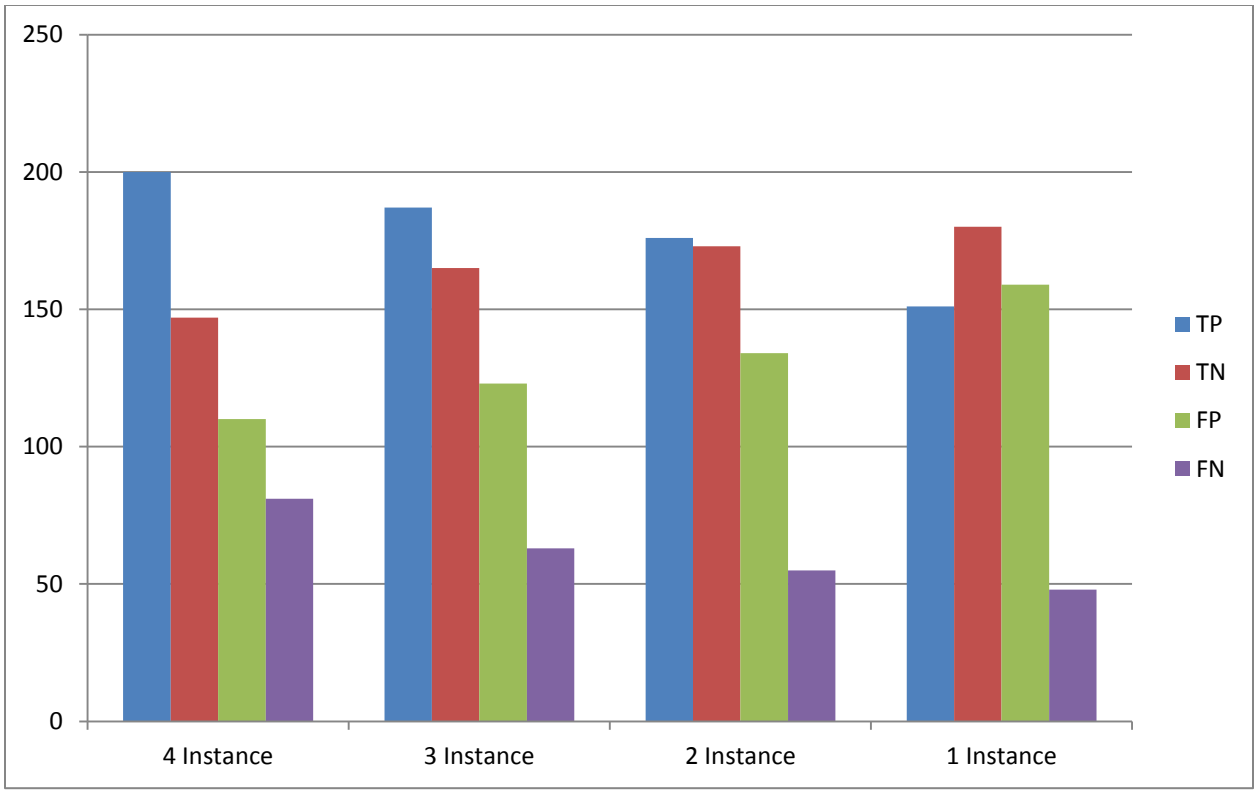

Figure B.1: Comparison of model predictions with genome-scale essentiality indices (EI) on minimal media across 4 libraries. Instances indicate the threshold of libraries for qualifying a gene as lethal. Positive results indicate predicted non-lethal genes, negative results indicate predicted lethal-genes. TP: true positive, model and EI both predict non-lethality; TN: true negative, model and EI both predict lethality; FP: false positive, model predicts non-lethality, EI predicts lethality; FN: false negative, model predicts lethality, EI predicts non-lethality.

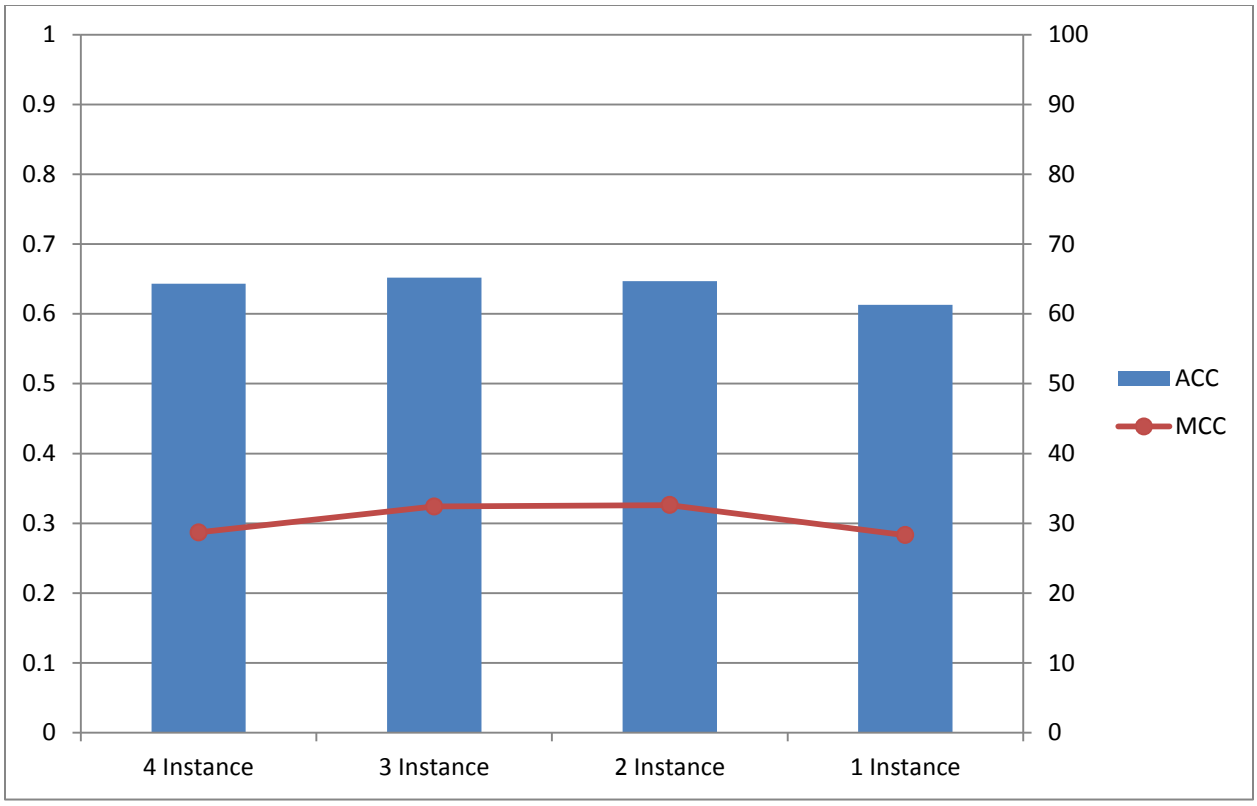

Figure B.2: A comparison of model predictions with genome-scale essentiality indices (EI) on minimal media across 4 libraries. We used Matthews Correlation Coefficient (MCC; left y-axis) and predictive accuracy (ACC; right y-axis) to compare our model's predictions against the EI predictions. Instances indicate the threshold of libraries for qualifying a gene as lethal.

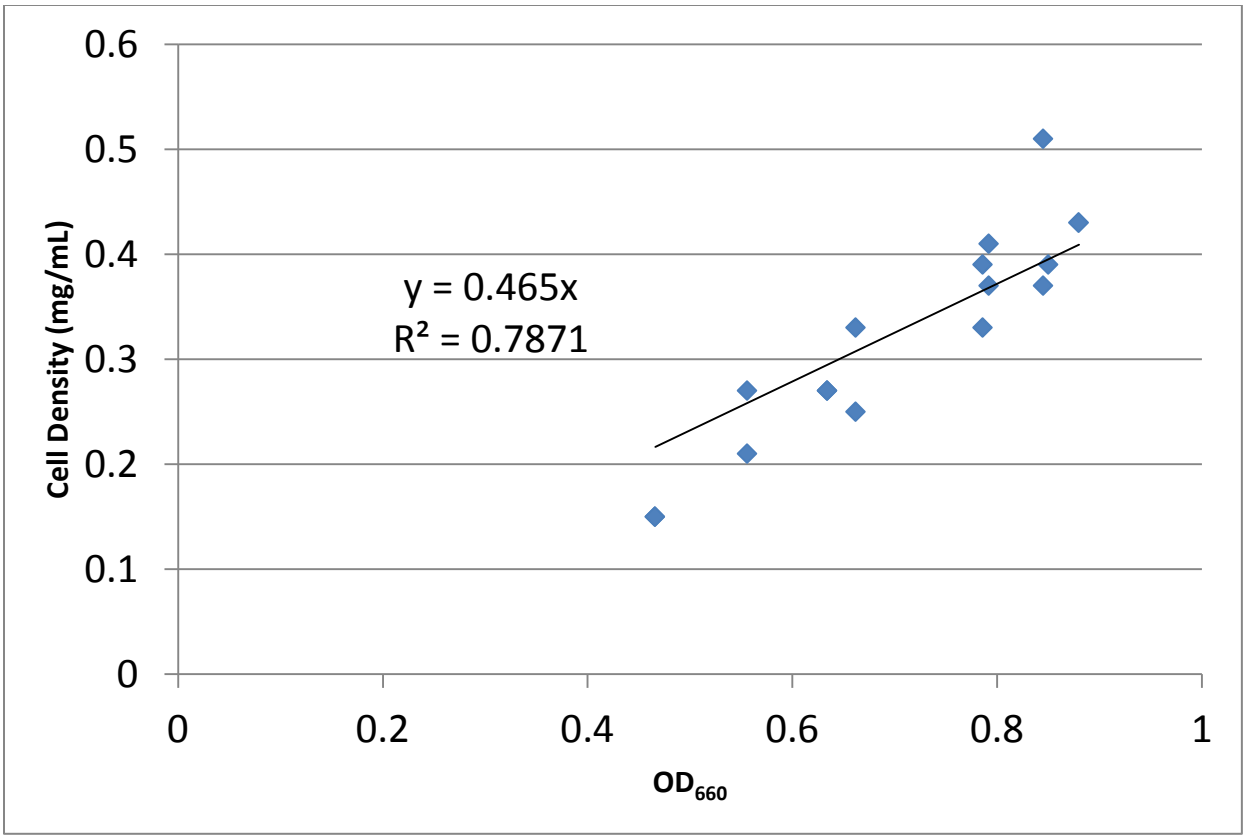

Figure B.3: Determination of the relationship between cell density and optical density (OD_{660}). Linear regression was set to intersect (0,0), as cell density must necessarily be 0 when $OD_{660} = 0$. For specific methodology on how these points were gathered, see Chapter 3 Methods.

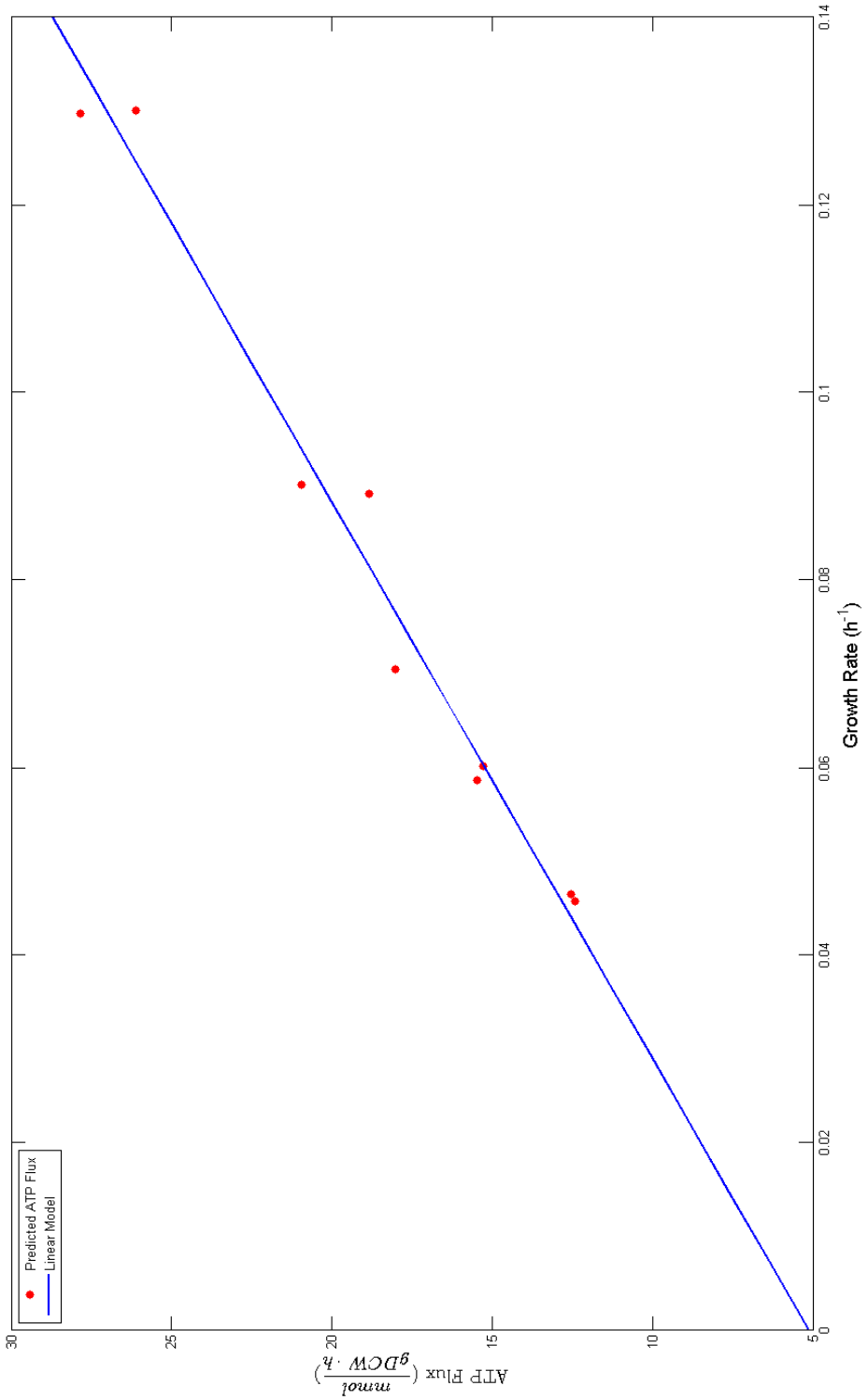

Figure B.4: Illustration of the process used to determine ATP maintenance values (see Chapter 3 Methods). Using all 9 measured samples, GAM (slope) and NGAM (y-intercept) were determined as 168.4 and 5.12 (mmol per grams [cell mass] per hour), respectively.

Supplementary File B.1: Example scripts for simulating iMR540. Reproduced here are the following scripts: "optimizeThermoModel" code that adds overall free energy estimation to an FBA solution; "maxGrowthOnH2" code that simulates maximum growth on H_2+CO_2 and prints out relevant fluxes; "simulateKOPanel" code that creates the knockout validation panel shown in Chapter 3. A full set of scripts for working with iMR540 can be found on GitHub (https://github.com/marichards/methanococcus)

```
function [solution,gibbs_flux,model] =
optimizeThermoModel(model,substrateRxns,concentrations,T,water_rxn,constraint_flag)

%%
% Version 4: 06/23/2015
%
% Accepts a model and necessary parameters for estimating thermodynamics of
% the overall system. Adds a new metabolite "dG" to the system that measures
% the free energy contribution for each exchange metabolite and a new
% reaction "GIBBS_kJ/GDW" that sums overall free energy for the system
%
% INPUT:
% model: a COBRA Toolbox model structure with a freeEnergy field
% substrateRxns: a set of exchange reactions for metabolites with specified
% concentrations
% concentrations: a set of concentrations corresponding to the specified
% substrateRxns (in mM)
% T: temperature (in Kelvin) for simulating growth
% water_rxn: identity of the water exchange reaction in the model. Water is
% treated separately from the aqueous metabolites and must be specified
% here.
%
% OUTPUT:
% solution: an FBA flux distribution that optimizes the supplied model
% gibbs_flux: an estimation of free energy produced by the model in the
% specified flux distribution (in kJ/gDCW/h)
% model: the supplied model with the addition of an overall reaction,
% GIBBS_kJ/GDW, that sums free energy for exchange reactions flowing in and
% out of the model
%
% Matthew Richards, 10/06/2015

% dG values are at pH=7.0 and ionic strength of 0.1 M

% Don't constrain solutions to be negative if there's no flag
if nargin < 6
    constraint_flag = false;
end

% Catch concentrations that are 0
if any(~concentrations)
    solution = optimizeCbModel(model,[],'one');
    gibbs_flux = inf;
```

Supplementary File B.1 (cont.)

```
else

    % Give an error for things not the same size
    if length(substrateRxns) ~= length(concentrations)
        error('substrateRxns and concentrations must be of equal length')
    end

    % Gas constant specification
    R = 8.314e-6; %kJ/mmol*K

    % Add the new reaction first, which adds the metabolite
    model = addReaction(model,'GIBBS_kJ/GDW','dG <=> ');
    % Give it no free energy of its own
    [~,gibbs_idx] = intersect(model.rxns,'GIBBS_kJ/GDW');
    model.freeEnergy(gibbs_idx) = 0;
    % Find index of dG
    [~,met_idx] = intersect(model.mets,'dG');

    % Alter the free energy values for things with substrate reactions in
    % the free energy vector itself
    % First grab the index of the exchange reactions in the model
    [rxns,rxn_idx] = intersect(model.rxns,substrateRxns,'stable');
    % Make a dictionary
    dict = containers.Map(rxns,rxn_idx);
    % For those indices, change the free energy numbers using concentration
    % Loop: put in the correct free energy term for each:
    % Basis: dG = dG_0 + RTln(C)
    for i = 1:length(substrateRxns)
       % Change the dG weight for the exchange reaction (Conc in mM)
       model.freeEnergy(dict(substrateRxns{i})) = model.freeEnergy(dict(substrateRxns{i}))...
           +R*T*log(concentrations(i));
    end

    % Add free energy values to S matrix for every one at once
    model.S(met_idx,:) = model.freeEnergy;

    % Go back and fix the Gibbs reaction to take in dG
    model.S(met_idx,gibbs_idx) = -1;

    % New Part (4/30/2013)
    % Add water contribution, which isn't reflected elsewhere
    [~,rxn_idx] = intersect(model.rxns,water_rxn);
    model.S(met_idx,rxn_idx) = model.freeEnergy(rxn_idx);

    % Let the free energy be as low as it desires
    model = changeRxnBounds(model,'GIBBS_kJ/GDW',-inf,'l');
```

Supplementary File B.1 (cont.)

```
    % If there's a constraint flag, then make free energy have to be
    % negative
    if constraint_flag
        model = changeRxnBounds(model,'GIBBS_kJ/GDW',0,'u');
    end

    % Simulate the model with minimization of overall flux
    solution = optimizeCbModel(model,[],'one');

    % Find the gibbs flux
    if isempty(solution.x)
        % If no solution, return that
        fprintf('\nNO THERMODYNAMICALLY FEASIBLE SOLUTION\n')
        gibbs_flux = [];
    else
        [~,idx] = intersect(model.rxns,'GIBBS_kJ/GDW');
        gibbs_flux = solution.x(idx);
    end
end
```

Supplementary File B.1 (cont.)

```
function [solution,gibbs_flux,model] =
maxGrowthOnH2(model,substrate_rxns,concentrations,print_flag)

% Simulate M. maripaludis growth on CO2 and H2 media, with ammonia as the
% nitrogen source. Print out the growth rate and relevant fluxes, return
% the full solution, the predicted free energy generation, and the modified
% model with the overall Gibbs free energy reaction added to the S matrix
%
% INPUT
% model: the M. maripaludis model, a COBRA Toolbox model structure
%
% OPTIONAL INPUT
% substrate_rxns: a list of exchange reactions in the M. maripaludis model
% for which a known concentration will be supplied. If supplied, it must be
% accompanied by a corresponding "concentrations" array. (Default =
% {'EX_cpd00011[e]','EX_cpd11640[e0]','EX_cpd01024[e0]'})
% concentrations: a list of effective concentrations in mM corresponding to
% the exchange reactions listed in "substrate_rxns". (Default = [1 1 1])
%
% OUTPUT
% solution: a flux distribution solution from running FBA on the M.
% maripaludis model that maximizes biomass yield
% gibbs_flux: model prediction of overall free energy generation, based on
% the model exchange fluxes in the solution
% model: the M. maripaludis model, with an additional reaction
% (GIBBS_kJ/GDW) that predicts overall free energy generation
%
% Matthew Richards, 09/24/2015

% Check if print_flag is supplied
if nargin < 4
    % Set default to true
    print_flag = true;
end

% Ensure that H2 is the electron source
model = switchToH2(model);

% Make sure that ammonia is the nitrogen source
model = switchToNH3(model);

% Specify substrate reactions and concentrations as 1 mM if not given
if nargin<2

    substrate_rxns = {'EX_cpd00011[e0]','EX_cpd11640[e0]','EX_cpd01024[e0]'};
    concentrations = [1 1 1];
```

Supplementary File B.1 (cont.)

```
    warning_flag = 1;
end
%Solve by maximizing biomass
[solution,gibbs_flux,model] = optimizeThermoModel(model,substrate_rxns...
    ,concentrations,310,'EX_cpd00001[e0]');

% Check for print flag
if print_flag

    %Pull out the overall reaction CO2 + 4H2 --> CH4 + 2H2O
    %Find the reaction indices
    [~,h2_idx]  = intersect(model.rxns,'EX_cpd11640[e0]');
    [~,co2_idx] = intersect(model.rxns,'EX_cpd00011[e0]');
    [~,ch4_idx] = intersect(model.rxns,'EX_cpd01024[e0]');
    [~,h2o_idx] = intersect(model.rxns,'EX_cpd00001[e0]');
    [~,form_idx] = intersect(model.rxns,'EX_cpd00047[e0]');
    [~,nh3_idx] = intersect(model.rxns,'EX_cpd00013[e0]');
    [~,po4_idx] = intersect(model.rxns,'EX_cpd00009[e0]');
    [~,ac_idx] = intersect(model.rxns,'EX_cpd00029[e0]');

    %Print the biomass flux
    fprintf('\n\nBiomass flux: %f\n\n',solution.f);
    %Print the reaction fluxes
    fprintf('Formate flux: %f\n',solution.x(form_idx))
    fprintf('CO2 flux: %f\n',solution.x(co2_idx))
    fprintf('H2 flux: %f\n',solution.x(h2_idx))
    fprintf('H2O flux: %f\n',solution.x(h2o_idx))
    fprintf('CH4 flux: %f\n',solution.x(ch4_idx))
    fprintf('NH3 flux: %f\n',solution.x(nh3_idx))
    fprintf('PO4 flux: %f\n',solution.x(po4_idx))
    fprintf('Acetate flux: %f\n',solution.x(ac_idx))

    %Print the per-CO2 actual reaction
    fprintf('\nOverall reaction:\nCO2 + 4 H2 --> 2 H2O + CH4\n')
    fprintf('\nModel overall reaction (per mole CH4)\n')
    fprintf('%0.2f CO2 + %0.2f H2 --> %0.2f H2O + CH4\n\n',-
solution.x(co2_idx)/solution.x(ch4_idx),...
        -solution.x(h2_idx)/solution.x(ch4_idx),solution.x(h2o_idx)/solution.x(ch4_idx))

    %Print the yield coefficient (grams biomass per mole CH4 produced)
    fprintf('Measured Yield Coefficient: 4.11 +/- 0.83 gDCW/mol CH4\n')
    fprintf('Predicted Yield Coefficient: %0.2f gDCW/mol
CH4\n\n',solution.f*1000/solution.x(ch4_idx)/log(2))

    %Find the ATP reaction index
    [~,atp_idx] = intersect(model.rxns,'ATPS');
```

Supplementary File B.1 (cont.)

```
    %Print the ATP yield coefficient (ATP per CH4)
    fprintf('Expected ATP/CH4 Yield: 0.5\n')
    fprintf('Predicted ATP/CH4 Yield: %0.3f\n\n', solution.x(atp_idx)/solution.x(ch4_idx))
end


%Add a warning for simulations with no concentrations given
if warning_flag
    warning('All external metabolite concentrations set to 1 mM');
end

% Print out the gibbs free energy prediction
if print_flag

    fprintf('Predicted Free Energy Generation: %f kJ/gDCW\n\n',gibbs_flux)

end
```

Supplementary File B.1 (cont.)

```
function simulateKOPanel(model)

% For the M. maripaludis S2 model, simulate the model for known gene KO
% experiments to get predictions and compare predictions to reality. Do all
% possible KOs for all 4 conditions, not just the replicates of experiments
%
% INPUT
% model: the M. maripaludis model, a COBRA Toolbox model structure
%
% Matthew Richards, 09/24/2015


% Alteration on 05/26/2015: Add MCC and accuracy calculations
% Create TP/TN/FP/FN metrics to fill up on appropriate things
tp = 0; tn = 0; fp = 0; fn = 0;

% Set GAPOR off to begin with
model = changeRxnBounds(model,'rxn07191[c0]',0,'b');
% Make sure model is set to H2
model = switchToH2(model);
% Set methane and EhA/Ehb Bounds on model
model = setMethaneSecretion(model,50);

% H2-CO2 simulations
fprintf('==============================\nGrowth on H2 +
CO2\n==============================');

% First simulate Wild-type growth
solution = optimizeCbModel(model,[],'one');
fprintf('\nWild-Type Growth: %0.2f\n\n',solution.f);
wt_growth = solution.f;

% Simulate Hmd KO (mmp0127)
ko_model = deleteModelGenes(model,'mmp0127',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-Hmd Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end
```

Supplementary File B.1 (cont.)

```
% Simulate Mtd KO (mmp0372)
ko_model = deleteModelGenes(model,'mmp0372',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-Mtd Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate FrcA KO (mmp0820)
ko_model = deleteModelGenes(model,'mmp0820',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FrcA Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate FruA KO (mmp1382)
ko_model = deleteModelGenes(model,'mmp1382',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FruA Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate FrcA-FruA double KO (mmp0820 and mmp1382)
ko_model = deleteModelGenes(model,{'mmp0820','mmp1382'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FrcA-FruA Growth Ratio: %0.2f\n',solution.f/wt_growth);
```

Supplementary File B.1 (cont.)

```
% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate  VhuAU-VhcA triple KO (mmp1694, mmp1693, mmp0823)
ko_model = deleteModelGenes(model,{'mmp0680','mmp1694','mmp1693','mmp0823'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-VhuAU-VhcA Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate  HdrB2 KO (mmp1053)
ko_model = deleteModelGenes(model,{'mmp0680','mmp1053'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-HdrB2 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate  FdhA1 KO (mmp1298)
ko_model = deleteModelGenes(model,{'mmp1298'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FdhA1 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate  FdhA2 KO (mmp0138)
ko_model = deleteModelGenes(model,{'mmp0138'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FdhA2 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate  FdhA1-FdhA2 double KO (mmp1298 and mmp0138)
ko_model = deleteModelGenes(model,{'mmp1298','mmp0138'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FdhA1-FdhA2 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate  FdhA2B2 KO (mmp0138 and mmp0139)
ko_model = deleteModelGenes(model,{'mmp0680','mmp0138','mmp0139'},0);
```

Supplementary File B.1 (cont.)

```
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FdhA2B2 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate EhbF KO (mmp1628)
ko_model = deleteModelGenes(model,'mmp1628',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-EhbF Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate 3H2ase KOs of frcAGB,fruAGB,hmd
% (mmp0820, mmp0818, mmp817, mmp1382, mmp1384, mmp1385, and mmp0127)
ko_model = deleteModelGenes(model,...
    {'mmp0680','mmp0820','mmp0818','mmp0817','mmp1382','mmp1384','mmp1385','mmp0127'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-3H2ase Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate 5H2ase KOs of frcA,fruA,hmd,vhuAU,vhcA
% (mmp0820, mmp1382,mmp0127,mmp1694,mmp1693,mmp0823)
ko_model = deleteModelGenes(model,...
    {'mmp0680','mmp0820','mmp1382','mmp0127','mmp1694','mmp1693','mmp0823'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-5H2ase Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (lethal)
if solution.f/wt_growth >= 0.1
    % Then add to False Positive
    fp = fp+1;
```

Supplementary File B.1 (cont.)

```
else
    % Then add to False Negative
    tn = tn+1;
end

% Simulate 6H2ase KOs of frcA,fruA,hmd,vhuAU,vhcA,ehbN
% (mmp0820, mmp1382, mmp0127,mmp1694,mmp1693,mmp0823,mmp1153)
ko_model = deleteModelGenes(model,...
    {'mmp0680','mmp0820','mmp1382','mmp0127','mmp1694','mmp1693','mmp0823','mmp1153'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-6H2ase Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (lethal)
if solution.f/wt_growth >= 0.1
    % Then add to False Positive
    fp = fp+1;
else
    % Then add to False Negative
    tn = tn+1;
end

% Simulate 6H2ase-cdh KOs of frcAGB,fruAGB,hmd,vhuAU,vhcA,ehbN, and cdh WITH CO supp
% (mmp0820, mmp0818, mmp817, mmp1382, mmp1384, mmp1385,
mmp0127,mmp1694,mmp1693,mmp0823,mmp1153,mmp0983-0995)
ko_model = deleteModelGenes(ko_model,...
    {'mmp0983','mmp0984','mmp0985'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-6H2ase-cdh Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Turn on GAPOR
model = changeRxnBounds(model,'rxn07191[c0]',-1000,'l');
model = changeRxnBounds(model,'rxn07191[c0]',1000,'u');

% Simulate 6H2ase_supp KOs of frcAGB,fruAGB,hmd,vhuAU,vhcA,ehbN
% (mmp0820, mmp1382, mmp0127,mmp1694,mmp1693,mmp0823,mmp1153)
ko_model = deleteModelGenes(model,...
    {'mmp0680','mmp0820','mmp1382','mmp0127','mmp1694','mmp1693','mmp0823','mmp1153'}...
    ,0);

solution = optimizeCbModel(ko_model,[],'one');
fprintf('-6H2ase_supp Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate 7H2ase_supp KOs of frcAGB,fruAGB,hmd,vhuAU,vhcA,ehbN,ehaNO
% (mmp0820, mmp1382, mmp0127,mmp1694,mmp1693,mmp0823,mmp1153,mmp1461,mmp1462)
ko_model = deleteModelGenes(model,...
```

Supplementary File B.1 (cont.)

```
{'mmp0680','mmp0820','mmp1382','mmp0127','mmp1694','mmp1693','mmp0823','mmp1153','mmp1461','m
mp1462'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-7H2ase Growth Ratio: %0.2f\n',solution.f/wt_growth);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Formate simulations
fprintf('\n================================\nGrowth on
Formate\n================================');
model = switchToFormate(model);

% Set GAPOR off to begin with
model = changeRxnBounds(model,'rxn07191[c0]',0,'b');

% First simulate Wild-type growth
solution = optimizeCbModel(model,[],'one');
fprintf('\nWild-Type Growth: %0.2f\n\n',solution.f);
wt_growth = solution.f;

% Simulate Hmd KO (mmp0127)
ko_model = deleteModelGenes(model,'mmp0127',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-Hmd Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate Mtd KO (mmp0372)
ko_model = deleteModelGenes(model,'mmp0372',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-Mtd Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
```

Supplementary File B.1 (cont.)

```
    % Then add to False Negative
    fn = fn+1;
end

% Simulate FrcA KO (mmp0820)
ko_model = deleteModelGenes(model,'mmp0820',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FrcA Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate FruA KO (mmp1382)
ko_model = deleteModelGenes(model,'mmp1382',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FruA Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate FrcA-FruA double KO (mmp0820 and mmp1382)
ko_model = deleteModelGenes(model,{'mmp0820','mmp1382'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FrcA-FruA Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end
```

Supplementary File B.1 (cont.)

```
% Simulate  VhuAU-VhcA triple KO (mmp1694, mmp1693, mmp0823)
ko_model = deleteModelGenes(model,{'mmp0680','mmp1694','mmp1693','mmp0823'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-VhuAU-VhcA Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate  HdrB2 KO (mmp1053)
ko_model = deleteModelGenes(model,{'mmp0680','mmp1053'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-HdrB2 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate  FdhA1 KO (mmp1298)
ko_model = deleteModelGenes(model,{'mmp1298'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FdhA1 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate  FdhA2 KO (mmp0138)
ko_model = deleteModelGenes(model,{'mmp0138'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FdhA2 Growth Ratio: %0.2f\n',solution.f/wt_growth);
```

Supplementary File B.1 (cont.)

```
% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate  FdhA1-FdhA2 double KO (mmp1298 and mmp0138)
ko_model = deleteModelGenes(model,{'mmp1298','mmp0138'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FdhA1-FdhA2 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (lethal)
if solution.f/wt_growth >= 0.1
    % Then add to False Positive
    fp = fp+1;
else
    % Then add to False Negative
    tn = tn+1;
end

% Simulate  FdhA2B2 KO (mmp0138 and mmp0139)
ko_model = deleteModelGenes(model,{'mmp0680','mmp0138','mmp0139'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FdhA2B2 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate EhbF KO (mmp1628)
ko_model = deleteModelGenes(model,'mmp1628',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-EhbF Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate 3H2ase KOs of frcAGB,fruAGB,hmd
% (mmp0820, mmp0818, mmp817, mmp1382, mmp1384, mmp1385, and mmp0127)
ko_model = deleteModelGenes(model,...
```

Supplementary File B.1 (cont.)

```
    {'mmp0680','mmp0820','mmp0818','mmp0817','mmp1382','mmp1384','mmp1385','mmp0127'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-3H2ase Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate 5H2ase KOs of frcA,fruA,hmd,vhuAU,vhcA
% (mmp0820, mmp1382,mmp0127,mmp1694,mmp1693,mmp0823)
ko_model = deleteModelGenes(model,...
    {'mmp0680','mmp0820','mmp1382','mmp0127','mmp1694','mmp1693','mmp0823'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-5H2ase Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (lethal)
if solution.f/wt_growth >= 0.1
    % Then add to False Positive
    fp = fp+1;
else
    % Then add to True Negative
    tn = tn+1;
end

% Simulate 6H2ase KOs of frcAGB,fruAGB,hmd,vhuAU,vhcA,ehbN
% (mmp0820, mmp1382, mmp0127,mmp1694,mmp1693,mmp0823,mmp1153)
ko_model = deleteModelGenes(model,...
    {'mmp0680','mmp0820','mmp1382','mmp0127','mmp1694','mmp1693','mmp0823','mmp1153'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-6H2ase Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (lethal)
if solution.f/wt_growth >= 0.1
    % Then add to False Positive
    fp = fp+1;
else
    % Then add to True Negative
    tn = tn+1;
```

Supplementary File B.1 (cont.)

```
end

% Simulate 6H2ase-cdh KOs of frcAGB,fruAGB,hmd,vhuAU,vhcA,ehbN, and cdh WITH CO supp
% (mmp0820, mmp0818, mmp817, mmp1382, mmp1384, mmp1385,
mmp0127,mmp1694,mmp1693,mmp0823,mmp1153,mmp0983-0995)
ko_model = deleteModelGenes(ko_model,...
    {'mmp0983','mmp0984','mmp0985'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-6H2ase-cdh Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Turn on GAPOR
model = changeRxnBounds(model,'rxn07191[c0]',-1000,'l');
model = changeRxnBounds(model,'rxn07191[c0]',1000,'u');

% Simulate 6H2ase_supp KOs of frcAGB,fruAGB,hmd,vhuAU,vhcA,ehbN
% (mmp0820, mmp1382, mmp0127,mmp1694,mmp1693,mmp0823,mmp1153)
ko_model = deleteModelGenes(model,...
    {'mmp0680','mmp0820','mmp1382','mmp0127','mmp1694','mmp1693','mmp0823','mmp1153'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-6H2ase_supp Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate 7H2ase_supp KOs of frcAGB,fruAGB,hmd,vhuAU,vhcA,ehbN,ehaNO
% (mmp0820, mmp1382, mmp0127,mmp1694,mmp1693,mmp0823,mmp1153,mmp1461,mmp1462)
ko_model = deleteModelGenes(model,...

{'mmp0680','mmp0820','mmp1382','mmp0127','mmp1694','mmp1693','mmp0823','mmp1153','mmp1461','m
mp1462'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-7H2ase Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
```

Supplementary File B.1 (cont.)

```
    % Then add to False Negative
    fn = fn+1;
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Formate plus H2 simulations
fprintf('\n================================\nGrowth on Formate +
H2\n================================');
model = changeRxnBounds(model,'EX_cpd11640[e0]',-1000,'l');

% Set GAPOR off to begin with
model = changeRxnBounds(model,'rxn07191[c0]',0,'b');

% First simulate Wild-type growth
solution = optimizeCbModel(model,[],'one');
fprintf('\nWild-Type Growth: %0.2f\n\n',solution.f);
wt_growth = solution.f;

% Simulate Hmd KO (mmp0127)
ko_model = deleteModelGenes(model,'mmp0127',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-Hmd Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate Mtd KO (mmp0372)
ko_model = deleteModelGenes(model,'mmp0372',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-Mtd Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate FrcA KO (mmp0820)
ko_model = deleteModelGenes(model,'mmp0820',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FrcA Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate FruA KO (mmp1382)
ko_model = deleteModelGenes(model,'mmp1382',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FruA Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate FrcA-FruA double KO (mmp0820 and mmp1382)
ko_model = deleteModelGenes(model,{'mmp0820','mmp1382'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FrcA-FruA Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate  VhuAU-VhcA triple KO (mmp1694, mmp1693, mmp0823)
ko_model = deleteModelGenes(model,{'mmp0680','mmp1694','mmp1693','mmp0823'},0);
```

Supplementary File B.1 (cont.)

```
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-VhuAU-VhcA Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate  HdrB2 KO (mmp1053)
ko_model = deleteModelGenes(model,{'mmp0680','mmp1053'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-HdrB2 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate  FdhA1 KO (mmp1298)
ko_model = deleteModelGenes(model,{'mmp1298'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FdhA1 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate  FdhA2 KO (mmp0138)
ko_model = deleteModelGenes(model,{'mmp0138'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FdhA2 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate  FdhA1-FdhA2 double KO (mmp1298 and mmp0138)
ko_model = deleteModelGenes(model,{'mmp1298','mmp0138'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FdhA1-FdhA2 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate  FdhA2B2 KO (mmp0138 and mmp0139)
ko_model = deleteModelGenes(model,{'mmp0680','mmp0138','mmp0139'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FdhA2B2 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate EhbF KO (mmp1628)
ko_model = deleteModelGenes(model,'mmp1628',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-EhbF Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate 3H2ase KOs of frcAGB,fruAGB,hmd
% (mmp0820, mmp0818, mmp817, mmp1382, mmp1384, mmp1385, and mmp0127)
ko_model = deleteModelGenes(model,...
    {'mmp0680','mmp0820','mmp0818','mmp0817','mmp1382','mmp1384','mmp1385','mmp0127'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-3H2ase Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate 5H2ase KOs of frcA,fruA,hmd,vhuAU,vhcA
% (mmp0820, mmp1382,mmp0127,mmp1694,mmp1693,mmp0823)
ko_model = deleteModelGenes(model,...
    {'mmp0680','mmp0820','mmp1382','mmp0127','mmp1694','mmp1693','mmp0823'}...
    ,0);
```

Supplementary File B.1 (cont.)

```
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-5H2ase Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate 6H2ase KOs of frcAGB,fruAGB,hmd,vhuAU,vhcA,ehbN
% (mmp0820, mmp1382, mmp0127,mmp1694,mmp1693,mmp0823,mmp1153)
ko_model = deleteModelGenes(model,...
    {'mmp0680','mmp0820','mmp1382','mmp0127','mmp1694','mmp1693','mmp0823','mmp1153'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-6H2ase Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate 6H2ase-cdh KOs of frcAGB,fruAGB,hmd,vhuAU,vhcA,ehbN, and cdh WITH CO supp
% (mmp0820, mmp0818, mmp817, mmp1382, mmp1384, mmp1385,
mmp0127,mmp1694,mmp1693,mmp0823,mmp1153,mmp0983-0995)
ko_model = deleteModelGenes(ko_model,...
    {'mmp0983','mmp0984','mmp0985'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-6H2ase-cdh Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Turn on GAPOR
model = changeRxnBounds(model,'rxn07191[c0]',-1000,'l');
model = changeRxnBounds(model,'rxn07191[c0]',1000,'u');

% Simulate 6H2ase_supp KOs of frcAGB,fruAGB,hmd,vhuAU,vhcA,ehbN
% (mmp0820, mmp1382, mmp0127,mmp1694,mmp1693,mmp0823,mmp1153)
ko_model = deleteModelGenes(model,...
    {'mmp0680','mmp0820','mmp1382','mmp0127','mmp1694','mmp1693','mmp0823','mmp1153'}...
    ,0);
```

Supplementary File B.1 (cont.)

```
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-6H2ase_supp Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate 7H2ase_supp KOs of frcAGB,fruAGB,hmd,vhuAU,vhcA,ehbN,ehaNO
% (mmp0820, mmp1382, mmp0127,mmp1694,mmp1693,mmp0823,mmp1153,mmp1461,mmp1462)
ko_model = deleteModelGenes(model,...

{'mmp0680','mmp0820','mmp1382','mmp0127','mmp1694','mmp1693','mmp0823','mmp1153','mmp1461','m
mp1462'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-7H2ase Growth Ratio: %0.2f\n',solution.f/wt_growth);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Formate plus CO simulations
fprintf('\n================================\nGrowth on Formate +
CO\n================================');
model = changeRxnBounds(model,'EX_cpd11640[e0]',0,'l');
model = changeRxnBounds(model,'EX_cpd00204[e0]',-1000,'l');

% Set GAPOR off to begin with
model = changeRxnBounds(model,'rxn07191[c0]',0,'b');

% First simulate Wild-type growth
solution = optimizeCbModel(model,[],'one');
fprintf('\nWild-Type Growth: %0.2f\n\n',solution.f);
wt_growth = solution.f;

% Simulate Hmd KO (mmp0127)
ko_model = deleteModelGenes(model,'mmp0127',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-Hmd Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate Mtd KO (mmp0372)
ko_model = deleteModelGenes(model,'mmp0372',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-Mtd Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate FrcA KO (mmp0820)
ko_model = deleteModelGenes(model,'mmp0820',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FrcA Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate FruA KO (mmp1382)
ko_model = deleteModelGenes(model,'mmp1382',0);
```

Supplementary File B.1 (cont.)

```
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FruA Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate FrcA-FruA double KO (mmp0820 and mmp1382)
ko_model = deleteModelGenes(model,{'mmp0820','mmp1382'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FrcA-FruA Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate  VhuAU-VhcA triple KO (mmp1694, mmp1693, mmp0823)
ko_model = deleteModelGenes(model,{'mmp0680','mmp1694','mmp1693','mmp0823'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-VhuAU-VhcA Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate  HdrB2 KO (mmp1053)
ko_model = deleteModelGenes(model,{'mmp0680','mmp1053'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-HdrB2 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate  FdhA1 KO (mmp1298)
ko_model = deleteModelGenes(model,{'mmp1298'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FdhA1 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate  FdhA2 KO (mmp0138)
ko_model = deleteModelGenes(model,{'mmp0138'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FdhA2 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate  FdhA1-FdhA2 double KO (mmp1298 and mmp0138)
ko_model = deleteModelGenes(model,{'mmp1298','mmp0138'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FdhA1-FdhA2 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate  FdhA2B2 KO (mmp0138 and mmp0139)
ko_model = deleteModelGenes(model,{'mmp0680','mmp0138','mmp0139'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-FdhA2B2 Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate EhbF KO (mmp1628)
ko_model = deleteModelGenes(model,'mmp1628',0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-EhbF Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate 3H2ase KOs of frcAGB,fruAGB,hmd
% (mmp0820, mmp0818, mmp817, mmp1382, mmp1384, mmp1385, and mmp0127)
ko_model = deleteModelGenes(model,...
```

Supplementary File B.1 (cont.)

```
    {'mmp0680','mmp0820','mmp0818','mmp0817','mmp1382','mmp1384','mmp1385','mmp0127'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-3H2ase Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate 5H2ase KOs of frcA,fruA,hmd,vhuAU,vhcA
% (mmp0820, mmp1382,mmp0127,mmp1694,mmp1693,mmp0823)
ko_model = deleteModelGenes(model,...
    {'mmp0680','mmp0820','mmp1382','mmp0127','mmp1694','mmp1693','mmp0823'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-5H2ase Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate 6H2ase KOs of frcAGB,fruAGB,hmd,vhuAU,vhcA,ehbN
% (mmp0820, mmp1382, mmp0127,mmp1694,mmp1693,mmp0823,mmp1153)
ko_model = deleteModelGenes(model,...
    {'mmp0680','mmp0820','mmp1382','mmp0127','mmp1694','mmp1693','mmp0823','mmp1153'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-6H2ase Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (non-lethal)
if solution.f/wt_growth >= 0.1
    % Then add to True Positive
    tp = tp+1;
else
    % Then add to False Negative
    fn = fn+1;
end

% Simulate 6H2ase-cdh KOs of frcAGB,fruAGB,hmd,vhuAU,vhcA,ehbN, and cdh WITH CO supp
% (mmp0820, mmp0818, mmp817, mmp1382, mmp1384, mmp1385,
mmp0127,mmp1694,mmp1693,mmp0823,mmp1153,mmp0983-0995)
ko_model = deleteModelGenes(ko_model,...
    {'mmp0983','mmp0984','mmp0985'},0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-6H2ase-cdh Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Compare to experimental result (lethal)
if solution.f/wt_growth >= 0.1
    % Then add to False Positive
    fp = fp+1;
else
    % Then add to True Negative
    tn = tn+1;
end
```

Supplementary File B.1 (cont.)

```
% Turn on GAPOR
model = changeRxnBounds(model,'rxn07191[c0]',-1000,'l');
model = changeRxnBounds(model,'rxn07191[c0]',1000,'u');

% Simulate 6H2ase_supp KOs of frcAGB,fruAGB,hmd,vhuAU,vhcA,ehbN
% (mmp0820, mmp1382, mmp0127,mmp1694,mmp1693,mmp0823,mmp1153)
ko_model = deleteModelGenes(model,...
    {'mmp0680','mmp0820','mmp1382','mmp0127','mmp1694','mmp1693','mmp0823','mmp1153'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-6H2ase_supp Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Simulate 7H2ase_supp KOs of frcAGB,fruAGB,hmd,vhuAU,vhcA,ehbN,ehaNO
% (mmp0820, mmp1382, mmp0127,mmp1694,mmp1693,mmp0823,mmp1153,mmp1461,mmp1462)
ko_model = deleteModelGenes(model,...

{'mmp0680','mmp0820','mmp1382','mmp0127','mmp1694','mmp1693','mmp0823','mmp1153','mmp1461','m
mp1462'}...
    ,0);
solution = optimizeCbModel(ko_model,[],'one');
fprintf('-7H2ase Growth Ratio: %0.2f\n',solution.f/wt_growth);

% Addition on 5/26/2015: Add MCC and accuracy calculation
total = tp+tn+fp+fn;
% First calculate total accuracy
fprintf('\nGene Knockout Accuracy: %0.1f%%(%d/%d)\n',100*(tp+tn)/total,(tp+tn),total);

% Next, calculate MCC
mcc = (tp*tn-fp*fn)/sqrt((tp+fp)*(tp+fn)*(tn+fp)*(tn+fn));
fprintf('Matthews Correlation Coefficient: %0.2f\n',mcc)
```